GOODNESS-OF-FIT TEST FOR LARGE NUMBER OF SMALL DATA SETS

A Dissertation

by

HYUNEUI LEE

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Jeffrey D. Hart
Committee Members,	Ursula Müller-Harknett
	Huiyan Sang
	Ximing Wu
Head of Department,	Valen Johnson

August 2017

Major Subject: Statistics

ABSTRACT

A goodness-of-fit (gof) problem, i.e., testing whether observed data come from a specific distribution is one of the important problems in statistics, and various tests for checking distributional assumptions have been suggested. Most tests are for one data set with a large enough sample sizes. However, this research focuses on the gof problem when there are a large number of small data sets. In other words, we assume that the number of data sets p increases to infinity and the sample size of each small data set n is finite. In this dissertation, we will denote p and n as the number of data sets and the sample sizes of each data sets, respectively.

Since the primary interest of this dissertation is testing whether every small data set comes from a known parametric family of distributions with different parameters, it is important to choose a gof test invariant to parameters of unknown distribution. Hence, as a basic approach, we suggest applying empirical distribution function (edf) based gof tests to every small data set and then combining P-values to obtain a single test. Two P-value combining methods, moment based tests and smoothing based tests, are suggested and their pros and cons are discussed. Especially, the two moment based tests, Edgington's method and Fisher's method, are compared with respect to Pitman efficiency and asymptotic power. We also find conditions that guarantee that the asymptotic null distribution of moment based tests based on empirical P-values is the same as that based on exact P-values. When the null is a location and scale family, there is no difficulty in applying the suggested test procedures. However, when the null is not a location and scale family, edf-based tests may depend on unknown parameters. To handle such a problem, we suggest using unconditional P-values and this requires an additional step of estimating the distri-

bution of unknown parameters. Several issues related to estimating the distribution of unknown parameters and obtaining unconditional P-values are also discussed. The performance of suggested test procedures are investigated via simulations and these procedures are applied to microarray data.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professor Jeffrey D. Hart, Ursula Müller-Harknett and Huiyan Sang of the Department of Statistics and Professor Ximing Wu of the Department of Agricultural Economics.

The data analyzed for Chapter 4 was provided by Professor Robert S. Chapkin in Texas A&M University.

All work for the dissertation was completed by the student, under the advisement of Jeffrey D. Hart of the Department of Statistics.

Funding Sources

There are no outside funding contributions to acknowledge related to the research and compilation of this document.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

1.1 Statement of general problem

Many data sets in modern statistics have a large number of variables with small sample sizes. For example, gene expression data can have hundreds or thousands of genes with a low number of replications. When we have a large number of data sets with few replications, it can be crucial to know whether every small data set comes from a specific distribution, such as a normal distribution, because if we can verify that every small data set follows a normal distribution, we may use the standard t-test to perform tests about population means.

If we can combine all small data sets into a single data set, verifying distributional assumptions for a large number of small data sets turns into a simple problem, which is a classical goodness-of-fit (gof) problem. Such an approach, however, may not be relevant under some situations. For example, if every small data set comes from the same family of distributions but with different parameters, the approach is not appropriate. In this case, it is clear that checking distributional assumptions for a large number of small data sets is challenging, and this dissertation focuses on the problem.

In the dissertation, it is assumed that we have data of the form $X_i = (X_{i1}, \ldots, X_{in})$, $i = 1, \ldots, p$, where vectors are independent of each other and for each i, $X_{i1}, \ldots, X_{in}$, are independent observations from a density function f_i. We also assume that there are hundreds or thousands of data sets with few replications, such as 5 or 10. The primary interest is to test the null hypothesis $H_0 : f_i = f_0(\cdot|\theta_i)$, where $\mathcal{F} = \{f_0(\cdot|\theta) : \theta \in \Theta\}$ is a known parametric family of distributions, but $\theta_1, \ldots, \theta_p$ are unknown.

1.2 The basic approach

There may exist several viable solutions to the current problem. If $\theta_1, \ldots, \theta_p$ were known, one could apply empirical distribution function (edf) based tests to every small data set, testing in each case the null hypothesis that $F(X_{ij}|\theta_i)$ has a uniform distribution on the interval $[0, 1]$. However, $\theta_1, \ldots, \theta_p$ are not known, and this leads to difficulties that will be dealt with subsequently.

Another possible solution is to cluster data sets based on proximity between parameters and exploit the clustered data sets to apply the probability integral transformation or to estimate the density of residuals. This approach assumes that within a cluster X_{ij} are distributed as $F(\cdot|\hat{\theta}_k)$ and hence $F(X_{ij}|\hat{\theta}_k)$ follows the uniform distribution. Here, $\hat{\theta}_k$ is the parameter estimate of the k-th cluster. For example, if we are interested in testing whether all small data sets are normally distributed with different means μ_i, $i = 1, \ldots, p$, and the same standard deviation σ, we can aggregate many small data sets into a few data sets with large sample sizes. We can then test uniformity after applying the probability integral transformations or test normality of residuals. Such a method is problematic in the sense that it cannot guarantee a large sample size in each cluster. This may happen when just a few clusters do not suffice. Hence, the difficulties mentioned in the previous paragraph may happen again when the probability integral transformation is used. If one is testing normality, a problem is that the density of residuals from an alternative distribution may be very close to normality, as shown in Litton (2009). Another difficulty arises when the alternative is local, i.e., when only a few data sets have different distributions than those specified by H_0. Such *local* alternatives could be masked if the small data sets are grouped into clusters.

One approach which does not have difficulties as mentioned above is to apply a

gof test to every small data set and combine P-values. In this dissertation gof tests based on edf will be used. Let $\hat{\theta}_i$ be an estimate of θ_i based only on $X_{i1}, \ldots, X_{in}$. If the distribution of an edf-based test statistic does not depend on θ_i, as in the case of a location-scale family, then a straightforward method based on simulation can be used to produce p P-values that are approximately independent and identically distributed as the uniform distribution under the null hypothesis. Specific P-value combining methods and choice of edf-based gof tests will be discussed in Chapter 2.

1.3 Review of nonparametric goodness-of-fit tests

The gof problem, i.e., testing whether observed data come from a specific distribution, is one of the important and classical problems in statistics, since even simple statistical methodologies, such as the t-test, analysis of variance and linear discriminant analysis, assume that data come from normal distributions. Of course, if this distributional assumption is not satisfied, results obtained from statistical methodologies are not necessarily reliable. Hence, various tests for checking a distributional assumption have been suggested. Pearson's chi-squared test (Pearson, 1900) is a popular test which can be used to test whether data come from a given distribution. Also, Kyriakoussis et al. (1998) suggested gof tests for Poisson, binomial, and negative binomial distributions. Their test is based on the characteristics of the first two moments of distributions. There exist a variety of tests for continuous variables. The Shapiro-Wilk test (Shapiro and Wilk, 1965) is a test of normality based on the ratio of the square of a linear combination of order statistics to the usual variance estimate. There are tests based on kernel density estimation. For example, Fan (1994, 1998) suggested a gof test exploiting L_2 distance between a kernel density estimate and a specified null distribution. Cao and Lugosi (2005) proposed a gof test based on minimizing the L_1 distance between a kernel density estimate and densities

belonging to the hypothesized class. Also, Rudzkis and Bakshev (2013) introduced a test statistic based on the maximum of a normalized deviation of a kernel density estimate from its expected value with respect to a hypothesized distribution. Song (2002) suggested a test based on the Kullback-Leibler Information Criterion (KLIC). In this paper, Song uses a sample entropy estimator due to Vasicek (1976) to estimate the KLIC and derives the asymptotic distribution of the test statistic. The smooth test, which is based on the probability integral transformation, to detect a smooth departure from the null hypothesis was investigated in Inglot and Ledwina (2006), Kallenberg and Ledwina (1997), Ledwina (1994) and Rayner et al. (2009).

The aforementioned tests are for one data set and most of the tests need a large enough sample size to obtain good power. There exist a few articles that deal with the gof problem of a large number of small data sets. Liang et al. (2009) proposed a generalized Shapiro-Wilk test for high-dimensional normality by using the theory of spherical distributions. Cox and Solomon (1986) proposed graphical and formal procedures to detect departures from the assumption that many small samples are distributed as the standard normal distribution. Their results may be applied to test whether a large number of small data sets are drawn from normal distributions. However, their test cannot be used to check if a large number of small data sets come from distributions other than normal distributions. Zhan and Hart (2012) devised a test based on the distance between kernel density estimates from small data sets and the average of all density estimates to test equality of a large number of densities. The test proposed by Zhan and Hart (2012) has the limitation that the test cannot inform whether or not every small data set is drawn from a specific distribution.

1.4 Importance of using goodness-of-fit test invariant to parameters of unknown distribution

We shall refer to the test statistic applied to each small data set as T. The test that combines all the P-values corresponding to different applications of T is called $\mathcal{T}$. Selecting the type of test for T is important because it will affect power and computing time. One criterion that should be considered is whether a test is invariant to parameters of unknown distributions. If T is not invariant to parameters of the unknown distribution, then the distribution of parameter values from one small data set to another will need to be inferred to obtain P-values that are identically distributed as the uniform distribution under the null. This additional step, inferring the distribution of parameters, may cause two problems. One problem is computing time, and the other problem is possible losses in power or lack of control of the size of tests. If the estimated distribution of parameters is not close to the true one, we may lose power or obtain a size greater than the nominal significance level. It is clear that finding the distribution of unknown parameters is unnecessary if a test is invariant to parameters of the unknown distributions. Hence, using a test that is location and scale invariant is crucial, and edf-based gof tests have this desirable property. Also, these gof tests are easy to compute. Hence, in this dissertation, we will focus on edf-based gof tests.

This dissertation is organized as follows: in Chapter 2, test procedures when the null distribution is a location and scale family are proposed, and pros and cons of the procedures are discussed. The power of the suggested test is investigated via simulation in Chapter 3. In Chapter 4, the suggested test is applied to microarray data. A test procedure when the null distribution is not a location and scale family is suggested in Chapter 5. In the last chapter, we give a summary of the dissertation

and discuss possible future study.

2. METHODOLOGY FOR LOCATION AND SCALE FAMILY

It is assumed that we observe data of the form $X_i = (X_{i1}, \ldots, X_{in}), i = 1, \ldots, p$, where the vectors are independent of each other and for each i, $X_{i1}, \ldots, X_{in}$ are independent observations from a density function f_i. The primary interest is to test the null hypothesis $H_0 : f_i = f_0(\cdot|\theta_i)$, where $\mathcal{F} = \{f_0(\cdot|\theta) : \theta \in \Theta\}$ is a known parametric family of distributions, but $\theta_1, \ldots, \theta_p$ are unknown. In this chapter, it is assumed that the parametric family is a location and scale family.

2.1 Selection of Test Statistics

As we discussed in Section 1.4, we need to consider tests invariant to location and scale parameters. Of many statistics with this desirable property, we apply edf-based gof tests such as Kolmogorov-Smirnov test (KS), Anderson-Darling test (AD), Cramér-von Mises test (CvM) and Watson test (Watson, 1961) to each small data set and then use all P-values to obtain a single test of the null hypothesis. Among these tests, the Watson test was originally devised as a gof test on a circle. However, it can be applied to observations on the line, because the test statistic does not depend on the fact that observations are on a circle. There are two reasons that these edf-based gof tests may be preferred to many other location and scale free tests under current setting, i.e., a large number of data sets with small sample sizes. First, these methods are computationally efficient. The computational efficiency is crucial because the test statistics are required to be computed for every small data set. Second, these test statistics do not depend on parameters that must be arbitrarily chosen by the user. For example, the KLIC based test (Song, 2002) is also invariant to location and scale and its test statistic can be efficiently computed. To implement the KLIC test, however, it is necessary to choose the order of spacings to estimate

the entropy, which might affect the stability of test statistics. Hence, the edf-based test statistics seem to be the most appropriate to the current problem.

Of the four edf tests mentioned, AD, CvM and Watson will be considered. These three tests are selected due to their power. There exist modifications of AD and CvM to increase their power. For example, Green and Hegazy (1976) suggested modified KS, AD and CvM by using the fact that the mean of the i-th uniform order statistic is $i/(n+1)$, and they showed that there were power improvements over the usual KS, AD and CvM. However, these improvements were limited to some alternatives and sample sizes. Especially, when the sample sizes are small such as 5 or 10, the original AD and CvM tend to have better power than the modified tests. Since we deal with the gof test for a large number of small data sets, it would be enough to consider the usual AD or CvM.

There is research comparing the power of gof tests. For instance, Stephens (1974) showed that the power of AD, when testing composite normality, was comparable to that of the Shapiro-Wilk test (SW), which was primarily designed to test composite normality. Also, D'Agostino and Stephens (1986, p.110) recommend AD because a departure in the tails is often important to detect and AD is more powerful than CvM, when an alternative distribution departs from the null distribution in the tails. Also, Frain (2007) investigated power properties of six tests of normality, which are Pearson's chi-squared test, SW, AD, CvM, KS, and Jarque-Bera test (Jarque and Bera, 1980), when the alternative is an α-stable distribution. Except for two tests, SW and Jarque-Bera test, which are designed to test normality, AD has the best power and CvM usually has better power than KS. Sürücü (2008) compares the power of gof tests when testing whether data come from two-parameter exponential distributions. He compared four tests, Tiku test using the sample spacing (Tiku, 1980), AD, Shapiro-Wilk test for exponential distributions (Shapiro and Wilk, 1972)

and the correlation test (Filliben, 1975; Smith and Bain, 1976), and found that Tiku test and AD were considerably more powerful than the other two tests. Since means of order statistics are required to apply the Tiku test, AD has an advantage over the Tiku test in its computational simplicity. Arshad et al. (2002) showed that AD and CvM have better power than KS when testing whether data come from generalized Pareto distributions. Quesenberry and Miller (1977) considered seven tests including KS, AD, CvM, Pearson's chi-squared test, Greenwood (1946), modified Greenwood and Watson to test that data are from a uniform distribution on the unit interval and suggested that Watson was a good choice if one test were to be used exclusively. Also, Watson is expected to be more powerful than CvM when the alternative distribution is shifted in variance, because it has the form of a variance while CvM has the form of a second moment, as mentioned in Watson (1961). Gürtler and Henze (2000) suggested a gof test for Cauchy distributions based on the integrated L_2 distances between the empirical characteristic function and the characteristic function of the standard Cauchy distribution. Their simulation results show that three edf-based tests, KS, AD, and CvM have stable and comparable powers to the suggested test for some choice of weights. Since their test requires the integration and choice of unknown weight, edf-based tests seem to be still preferable.

When edf-based tests are applied to small data sets, a bias issue arises. In other words, their exist null distributions and alternatives such that the power of the test is smaller than the size of the test. Even if it is well known that AD, CvM, and Watson are consistent, this consistency is irrelevant to the current setting. For example, Massey (1950) and Thompson (1966) showed that KS and CvM are biased for certain sample sizes under some alternatives. Also, we can heuristically justify that the bias of the tests may depend on the shape of null and alternative distributions. Table 2.1 shows the power of three tests, AD, CvM and Watson, when we test whether

data come from the uniform distribution. Even though results of the size of three tests are not shown in the table, all three tests attain the right size. From now on, a beta distribution with parameters α and β will be denoted by beta(α, β). Two tests except Watson are biased when the data are from beta(2,2), but all tests are not biased when the data are from beta(1/2,1/2). These results accord with results from Quesenberry and Miller (1977). In their simulations, Watson is unbiased for all considered alternatives, while AD and CvM are biased for some alternatives. This suggests that Watson may be unbiased for more alternatives than AD or CvM, and this is another reason that Watson may be preferable to AD and CvM.

One possible explanation for the bias problem is the shape of distributions. Figure 2.1 shows the probability density function (pdf) and the cumulative distribution function (cdf) of a beta distribution. In the plot, we notice that both beta distributions show more departures from the uniform distribution around the tails. Also, we may expect more observations from the tails when the alternative is beta(1/2,1/2). This may explain the fact that tests are not biased when data are from beta(1/2,1/2). On the contrary, only a few observations are expected to be from the tails when data are from beta(2,2), implying difficulties to detect departures from the null. This simulation indicates the possibility that the test is biased, especially when we have data sets with small sample sizes, since testing whether data come from a fully specified distribution is equivalent to testing whether data come from the uniform distribution.

2.2 Methods of combining P-values and testing procedures

There are two ways to combine test results from small data sets. One is to use the test statistic itself, and the other is to use the P-value. The latter seems preferable since the distribution of the P-value under the null hypothesis is known whereas the distribution of test statistics under the null hypothesis is unknown. One possible

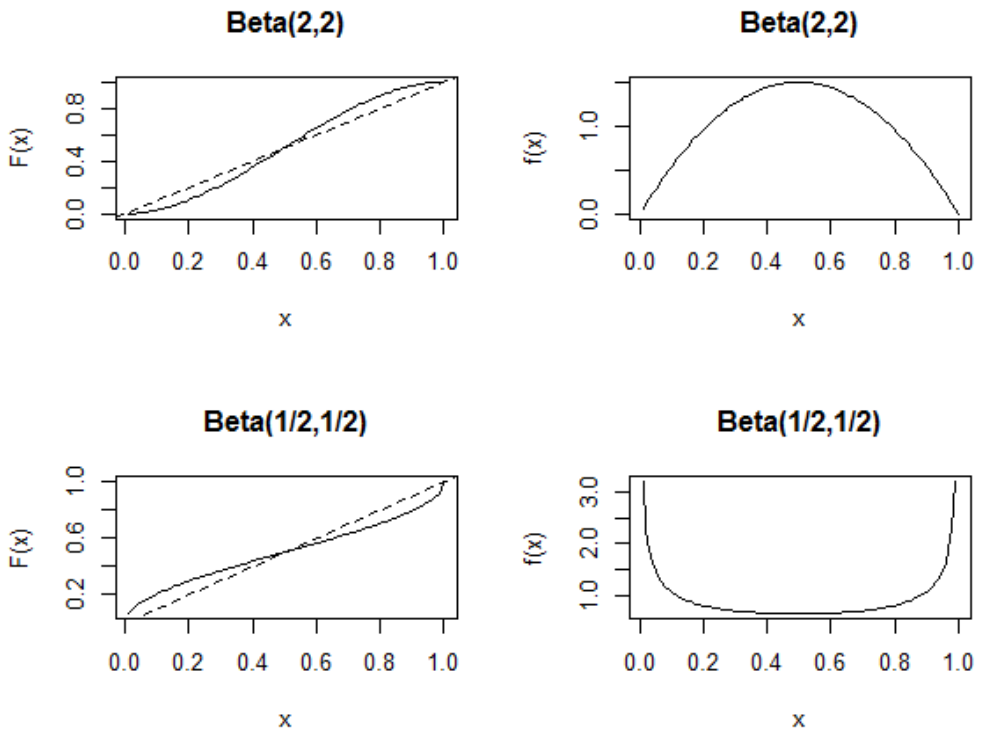

Figure 2.1: This figure shows that the probability density functions and the cumulative distribution functions of beta(1/2,1/2) and beta(2,2).

way to combine independent P-values is using methods like Fisher's method, the normal transformation method, the minimum P-value method, the maximum P-value method, the mean of P-value method, i.e., Edgington's method (Edgington, 1972), and the logit method (Mudholkar and George, 1977).

Of the possible combining methods, it seems best to choose one or two suitable methods. There exists much research comparing or evaluating parametric methods. For example, Birnbaum (1954) showed that if a combining method satisfies a general condition for admissibility, then we can find some alternative hypothesis for which the combining method gives the best test of the null hypothesis. The condition is

Table 2.1: This table shows the power(%) of three tests, AD, CvM and Watson. The numbers in parentheses are the means of P-values. Each number is obtained from 2,000 replications and the significance level α is 0.05.

	AD		CvM		Watson	
n	beta(1/2,1/2)	beta(2,2)	beta(1/2,1/2)	beta(2,2)	beta(1/2,1/2)	beta(2,2)
5	28.9	0.2	12.7	1.2	13.8	11.2
	(0.26)	(0.55)	(0.38)	(0.49)	(0.40)	(0.40)
10	36.3	1.0	12.4	2.3	23.9	20.0
	(0.20)	(0.46)	(0.33)	(0.43)	(0.30)	(0.30)

that if the null hypothesis is rejected for a vector $(P_1, \ldots, P_p)$ then it should also be rejected for a vector $(P_1^*, \ldots, P_p^*)$ such that $P_i^* \leq P_i$ for each i, where P_i is the P-value for the i-th data set. Lancaster (1961) developed a way to evaluate combining methods at a specified alternative distribution by representing the distribution in terms of orthonormal functions with respect to the null distribution. Littell and Folks (1971) showed that Fisher's method is asymptotically optimal in the sense of the Bahadur efficiency among four methods, which are Fisher's method, the normal transformation method, the minimum P-value method and the maximum P-value method. Berk and Cohen (1979) showed that the logit method is also asymptotically Bahadur optimal. Cohen et al. (1982) showed that the method of weighted sum of P-values has the same Bahadur slope as Fisher's method if and only if all tests have the same slope and the sample sizes for each data set satisfy the following condition: $n_i = \dfrac{n}{p} + o(n)$, where n_i is the sample size for the i-th data set, p is the number of combined tests, and $n = \sum_{j=1}^{p} n_j$. In their analysis, they assumed that n_i increases without bound and p is fixed. Loughin (2004) investigated the power of methods of combining P-values by simulation. In his simulation, distributions of the P-value are modeled by beta distributions with parameters $\alpha = 1$ and $\beta \geq 1$ because these distributions have appropriate properties, such as a non-increasing distribution with

a support between 0 and 1 and the possibility to control the strength of evidence against the null through β. He compared methods by using different numbers of tests, and different patterns and strengths of evidence against the null hypothesis in his simulations. The simulation results suggested that there is no uniformly most powerful method and the power of combining methods depends on the number of tests rejecting the null hypothesis and the strength of evidence against the null. For instance, if few of the data sets depart from the null and the strength of evidence is moderate, Fishers' method would be preferable. He does not recommend to use Edgington's method or the maximum P-value method because these two methods usually have very poor power. However, the simulation may not be valid if beta distributions with parameters $\alpha = 1$ and $\beta \geq 1$ are not an adequate model for the distribution of the P-value. Of course, the bias of tests may affect the shape of the distribution of the P-value, and its shape may be different from the expected non-increasing shape.

To investigate the effects of the bias on the shape of the distribution of the P-value, P-values, when testing uniformity, are obtained from 100 data sets of sample size 10 using CvM. To decrease the sampling variability, 100 iterations were used. The distribution of the P-value, in Figure 2.2, is estimated by the median of 100 kernel density estimates. For the beta(2,2) distribution, the P-value distribution is not non-increasing, indicating that the results from Loughin (2004) may not be relevant in the current problem. Thus, even though Loughin (2004) had not recommended the use of Edgington's method, we will still consider Edgington's method.

To select combining methods, we need some criterion. The evaluation method from Lancaster (1961) cannot be applied because the distribution of the P-value under the alternative hypothesis is not specified and we need to consider a general alternative hypothesis. Also, existing research on Bahadur efficiency may not be a

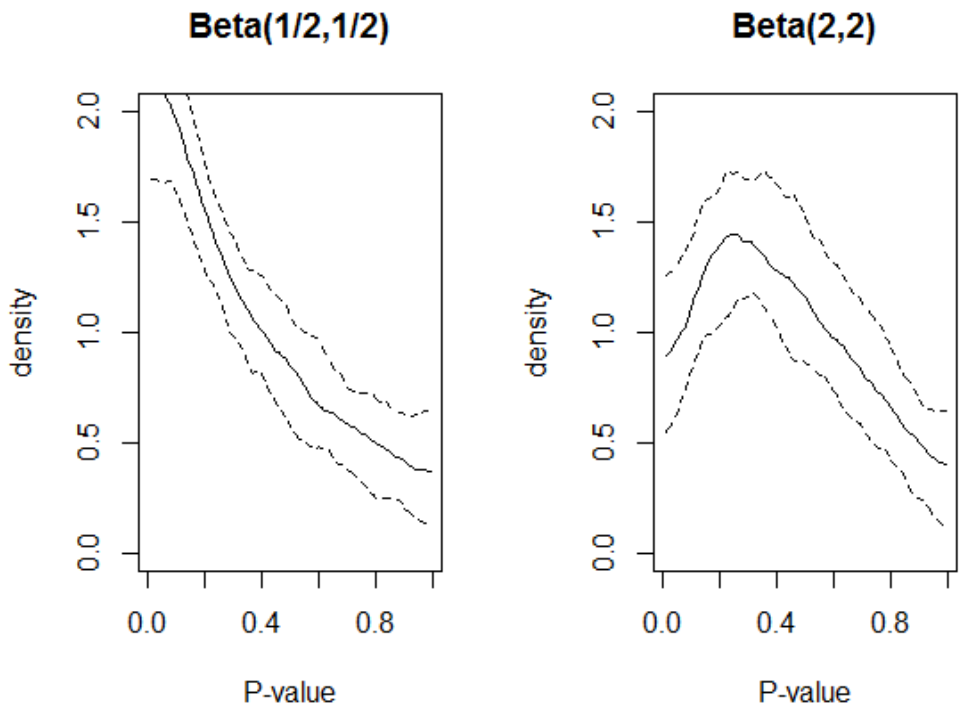

Figure 2.2: This figure shows density estimates of the P-value when the null hypothesis is false. The left and right plots are P-value distributions when the data come from beta(1/2,1/2) and beta(2,2), and CvM is used. The solid line is the median of kernel density estimates and dashed lines represent 0.025 and 0.975 percentiles of kernel density estimates.

reasonable way to select a method in the current setting, because Bahadur efficiency was computed under the assumption that the sample size for each data set increases to infinity and the number of combined tests is finite. In this dissertation, however, we need to combine tests whose number increases to infinity and the sample size for each small data set is finite. Fisher's method is selected because the simulation of Loughin (2004) showed that it detects evidence against the null especially well under local alternatives and it usually has at least 80% power under other patterns of alternative hypotheses. Also, this method satisfies the admissibility condition and

Table 2.2: This table shows the size(%) and power(%) of a nominal size 0.05 test when testing whether data come from the uniform distribution. The numbers in parentheses are local power when 90% of data sets come from the null distribution, i.e., the uniform distribution. The size and power are computed based on the one-sided critical value. CvM is used to compute the P-value, and each value is obtained from 2,000 replications.

		Edgington's method			Fisher's method		
n	p	uniform	beta(1/2,1/2)	beta(2,2)	uniform	beta(1/2,1/2)	beta(2,2)
5	100	5.65	98.6 (11.1)	5.3 (5.3)	5.1	99.2 (12.8)	0.1 (3.3)
	300	4.5	100.0 (15.2)	7.4 (4.8)	5.0	100.0 (18.3)	0.0 (3.3)
10	100	4.5	100.0 (14.7)	86.3 (7.7)	4.7	100.0 (16.0)	14.2 (4.9)
	300	5.9	100.0 (21.5)	100.0(11.4)	5.2	100.0 (26.2)	40.6 (6.7)

is more easily handled in a mathematical sense than is the normal transformation method, which is recommended for general use by Loughin (2004). Edgington's method is selected because it also satisfies the condition for admissibility and the method effectively detects evidence against the null when all null hypotheses are false (Edgington, 1972; Loughin, 2004). Also, the method can be handled easily in a mathematical sense like Fisher's method. Using the mathematical tractability of both methods, we will compare them asymptotically in the next section.

Before comparing the two methods, we will heuristically investigate their performance by a simple simulation when we have a finite number of data sets. The simulation results in Table 2.2 show power for both local and non-local alternatives. The local alternatives are such that 90% of data sets come from the null distribution and 10% from a non-null distribution. In the non-local alternatives, all data sets come from the same non-null distribution. For the non-local alternatives, Edgington's method has better power than Fisher's method. This result is related to the rejection regions of the combining methods. From Figure 2.3, we notice that Edgington's method may detect departures from the null better than Fisher's method when

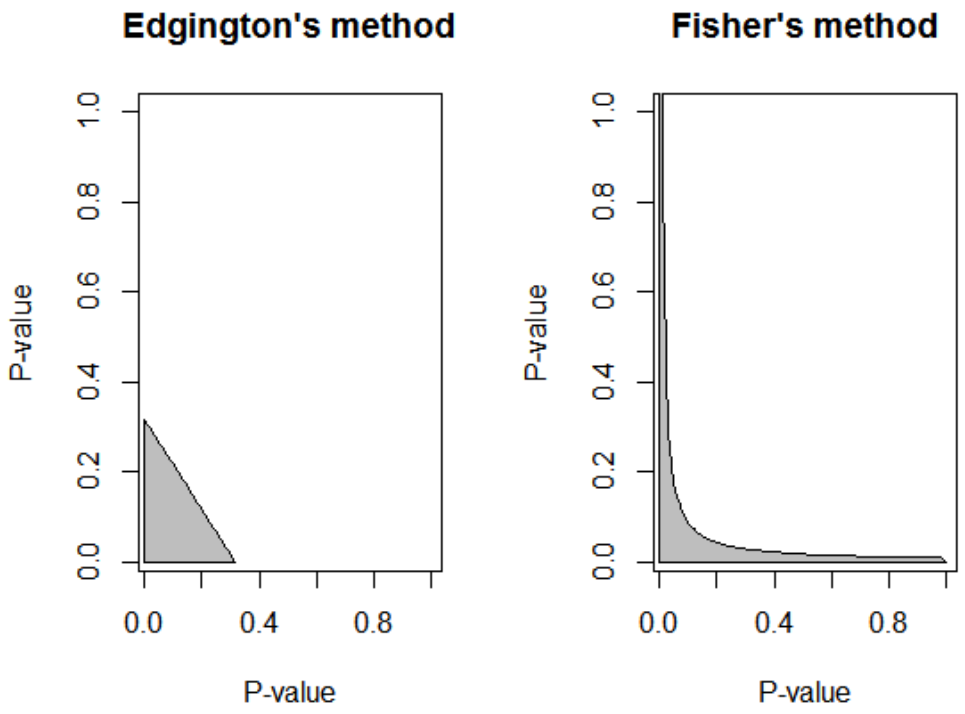

Figure 2.3: These plots show rejection regions corresponding to Edgington's method and Fisher's method when combining *two* P-values. The shaded area corresponds to rejection region of each method.

both small data sets come from the same non-null distribution. On the contrary, Fisher's method is expected to perform well when only one of the data sets comes from a non-null distribution because its rejection region in Figure 2.3 indicates that the null hypothesis can be rejected even if just one of the two P-values is close to 0. In addition to simulation results of Loughin (2004), these rejection regions and simulation results show that there does not exist a uniformly better method, and we may need to use both Fisher's method and Edgington's method. Hence, the testing procedure used will be as follows:

1. For every small data set, AD, CvM or Watson is applied.

2. By using the P-values, $P_1, \ldots, P_p$, $T_1 = -2\sum_{i=1}^{p} \log P_i$ and $T_2 = \dfrac{\sqrt{p}(\bar{P} - 1/2)}{\sqrt{1/12}}$ are computed.

3. The null hypothesis is rejected if $T_1 > \chi^2_{\alpha_1}(2p)$ or $T_1 < \chi^2_{1-\alpha_2}(2p)$, where $\chi^2_{\alpha}(2p)$ denotes the $(1-\alpha)$ percentile of the chi-squared distribution with d.f. $2p$ and $\alpha_1 + \alpha_2 = \alpha$. Similarly, the null hypothesis is rejected if $T_2 < Z_{\alpha_1}$ or $T_2 > Z_{1-\alpha_2}$, where Z_α denotes the $(1-\alpha)$ percentile of the standard normal distribution.

One important thing in the above test procedure is to apply two-sided tests. These are suggested because the tests may be biased when data sets have small sample sizes. The necessity of using two-sided tests after P-values are combined can be justified for each method. For Fisher's method, the method is uniformly most powerful when the distribution of the P-value is $F(z) = z^k$ where $0 < z < 1, k > 0$. It can be shown that a likelihood ratio test of uniformity under these alternatives is a two-sided test. If the distribution of the P-value is non-increasing, $0 < k < 1$, the null hypothesis of uniformity should be rejected when the test statistic is large. Similarly, if the distribution of the P-value is not non-increasing, $k > 1$, the uniformly most powerful test is to reject the null hypothesis of uniformity when the test statistic is small. This result along with the fact that the distribution of the P-value under alternatives does not have a non-increasing shape, especially when tests are biased, suggests the use of two-sided tests. For Edgington's method, the mean of the P-value may still be smaller than 0.5 even if the test is biased for some alternatives. However, Table 2.1 shows that the mean of the P-value can be greater than 0.5, especially when the test is biased.

The idea of using the two-sided tests in gof problems is not a new idea. Seshadri et al. (1969) compared the power of tests for exponential distributions based on two transformations, J and K, of observations to uniform observations under the

null hypothesis. They found that the J transformation is more likely to produce superuniform observations, more evenly spaced than uniform observations, when the original observations are from alternative distributions. Hence, they used the lower tail to detect a departure from the null distribution when they used the J transformation. Also, Fisher (1925, Section 20) wrote about the necessity of using the lower tail when doing Pearson's chi-squared test:

"The term Goodness of Fit has caused some to fall into the fallacy of believing that the higher the value of P the more satisfactorily is the hypothesis verified. Values over 0.999 have sometimes been reported which, if the hypothesis were true, would only occur once in a thousand trials. Generally, such cases are demonstrably due to the use of inaccurate formulae, but occasionally small values of χ^2 beyond the expected range do occur ... In these cases the hypothesis considered is as definitely disproved as if P had been 0.001."

Similarly, Yule and Kendall (1950, section 20.20.) wrote about the necessity of using the two-sided tests.

2.3 Comparison of Fisher's method and Edgington's method

In this section, two combining methods, Edgington's method and Fisher's method, are compared based on Pitman efficiency and the asymptotic power under $\sqrt{p}$-alternatives. Pitman efficiency is defined as the limiting ratio of the sample sizes required to obtain the same limiting power when a parameter under alternatives is different from what it is under the null by the amount of $O(1/\sqrt{p})$, where p is the sample size. For example, if the Pitman efficiency of a test S relative to a test T is 2, it implies that we would need approximately twice as many samples for the test T as for the test S to obtain the same asymptotic power. Since beta distributions have been used to model the distribution of the P-value (Allison et al., 2002; Loughin,

2004) and Parker and Rothenberg (1988) point out that any distribution on the interval $[0, 1]$ can be modeled as a mixture of beta distributions, Pitman efficiency is obtained under the assumption that the distribution of the P-value is a beta distribution with parameters $\alpha = 1$ and $\beta = \theta$. Since the distribution follows the uniform distribution when $\theta = 1$, the null and alternative hypotheses can be defined as

$$H_0 : \theta = 1 \text{ and } H_A : \theta = 1 + \frac{c}{\sqrt{p}} \tag{2.1}$$

To obtain Pitman effciency, we use the Noether theorem, which is stated in the appendix.

Proposition 2.3.1 *The Noether theorem (Randles and Wolfe, 1979, p.147) can be applied to tests that have upper-tailed rejection region and lower-tailed rejection region when they have asymptotic normal distributions under the null and alternatives.*

Proof This can be shown easily by using the fact that $\Phi(-x) = 1 - \Phi(x)$ where $\Phi(\cdot)$ is the distribution function of the standard normal distribution.

Proposition 2.3.2 *The Pitman efficiency of Edgington's method relative to Fisher's method under hypotheses (2.1) is 1.80 when tests are not biased.*

Proof To prove the proposition, we need to verify the conditions of Noether's theorem. Define θ_0 and θ_p to be the values of θ under the null and the alternative, respectively. Let $\mu_1(\theta) = \dfrac{1}{1+\theta}$ and $\mu_2(\theta) = -2(\psi(1) - \psi(1+\theta))$ where $\psi(\cdot)$ is the digamma function. Also, let $\sigma_1^2(\theta) = \dfrac{\theta}{p(1+\theta)^2(2+\theta)}$ and $\sigma_2^2(\theta) = \dfrac{4(\psi_1(1) - \psi_1(1+\theta))}{p}$, where $\psi_1(\cdot)$ is the trigamma function. Here, $\mu_1(\theta)$ and $\sigma_1(\theta)$ are the mean and standard deviation of beta$(1, \theta)$. Similarly, $\mu_2(\theta)$ and $\sigma_2(\theta)$ are the mean and standard deviation of $-2 \log X$, where X is distributed as beta$(1, \theta)$. For Edgington's method,

$\frac{\bar{P} - \mu_1(\theta)}{\sqrt{\sigma_1^2(\theta)}}$ converges to the standard normal distribution as p tends to ∞. Hence, the conditions A1 and A2 are verified. Also, $\frac{\sigma_1(\theta_p)}{\sigma_1(\theta_0)}$ converges to 1 as p goes to ∞, and by using the fact that $\mu_1^{'}(\theta) = -(1+\theta)^{-2}$, we can show that $\frac{\mu_1^{'}(\theta_p)}{\mu_1^{'}(\theta_0)}$ converges to 1 as p increases to ∞. Finally, we have

$$K_E := \lim_{p\to\infty} \frac{\mu_1^{'}(\theta_0)}{\sqrt{p\sigma_1^2(\theta_0)}} = \frac{-1/4}{\sqrt{1/12}}.$$

For Fisher's method, $\frac{\overline{-2\log P} - \mu_2(\theta)}{\sigma_2(\theta)}$ converges to the standard normal distribution as p increases to ∞. Since $\mu_2^{'}(\theta) = 2\psi_1(\theta)$, we can easily verify the conditions A4 and A5. Also, $\frac{\sigma_2(\theta_p)}{\sigma_2(\theta_0)}$ converges to 1 as p goes to ∞. Finally, we have

$$K_F := \lim_{p\to\infty} \frac{\mu_2^{'}(\theta_0)}{\sqrt{p\sigma_2^2(\theta_0)}} = \frac{2\psi_1(2)}{\sqrt{4(\psi_1(1) - \psi_1(2))}}.$$

Hence, the Pitman efficiency of Edgington's method relative to Fisher's method is $\frac{K_E^2}{K_F^2} = 1.80$.

Proposition 2.3.2 implies that about a 1.8 times larger sample size is required by Fisher's method to obtain the same power, indicating that Edgington's method is slightly better in the sense of Pitman efficiency. This might imply that Edgington's method is preferable to Fisher's method when there exists relatively weak evidence against the null in the sense of power.

Another criterion to compare the two methods is to investigate the asymptotic power. The asymptotic power is obtained under hypotheses (2.1) and another hypotheses defined as

$$H_0 : f = I_{(0,1)}(P) \text{ and } H_A : f = (1 - p^{-1/2})I_{(0,1)}(P) + p^{-1/2}g, \tag{2.2}$$

where $I_A(x) = \begin{cases} 0 & \text{if } x \notin A \\ 1 & \text{if } x \in A \end{cases}$ and g is a density function with support (0,1).

Proposition 2.3.3 *The asymptotic powers of Edgington's method and Fisher's method under hypotheses (2.1) are* $\Phi\left(-z_\alpha + \frac{c}{4\sqrt{1/12}}\right)$ *and* $1-\Phi(z_\alpha - c\psi_1(2))$*, respectively, where* $\psi_1(\cdot)$ *denotes the trigamma function.*

Proof For Edgington's method,

$$Pr\left(\frac{\sqrt{p}(\bar{P}-1/2)}{\sqrt{1/12}} < -z_\alpha\right)$$
$$= Pr\left(\frac{\sqrt{p}(\bar{P}-E_{H_A}(P))}{\sqrt{\mathrm{Var}_{H_A}(P)}} < \frac{\sqrt{1/12}}{\sqrt{\mathrm{Var}_{H_A}(P)}}\left(-z_\alpha - \frac{\sqrt{p}(E_{H_A}(P)-1/2)}{\sqrt{1/12}}\right)\right).$$

Under the considered alternatives, $\mathrm{Var}_{H_A}(P) = \frac{1+c/\sqrt{p}}{(2+c/\sqrt{p})^2(3+c/\sqrt{p})}$ and $E_{H_A}(P) = \frac{1}{2+c/\sqrt{p}}$. By using these facts, we can easily show $\frac{1/12}{\mathrm{Var}_{H_A}(P)} = 1 + o_p(1)$ and $\sqrt{p}(E_{H_A}(P) - 1/2) = -c/4 + o_p(1)$. This implies that the asymptotic power of Edgington's method under hypotheses (2.1) is equal $\Phi\left(-z_\alpha + \frac{c}{4\sqrt{1/12}}\right)$.

For Fisher's method,

$$Pr\left(\frac{\sqrt{p}(\overline{-2\log P}-2)}{\sqrt{4}} > z_\alpha\right)$$
$$= Pr\left(\frac{\sqrt{p}(\bar{X}-E_{H_A}(X))}{\sqrt{\mathrm{Var}_{H_A}(X)}} > \frac{\sqrt{4}}{\sqrt{\mathrm{Var}_{H_A}(X)}}\left(z_\alpha - \frac{\sqrt{p}(E_{H_A}(X)-2)}{\sqrt{4}}\right)\right),$$

where X denotes $-2\log P$.

Under considered alternatives,

$$E_{H_A}(-2\log P) = -2\left(\psi(1) - \psi\left(2+\frac{c}{\sqrt{p}}\right)\right)$$

$$= -2\left(\psi(1) - \psi(2) - \frac{c}{\sqrt{p}}\psi_1(2) + O\left(\frac{1}{p}\right)\right),$$

where the second equation holds by Taylor's expansion. Similarly,

$$\begin{aligned}\mathrm{Var}_{H_A}(-2\log P) &= 4\left(\psi_1(1) - \psi_1\left(2 + \frac{c}{\sqrt{p}}\right)\right) \\ &= 4\left(\psi_1(1) - \psi_1(2) - \frac{c}{\sqrt{p}}\psi_1'(2) + O\left(\frac{1}{p}\right)\right).\end{aligned}$$

By using these facts, the asymptotic power of Fisher's method is equal to $1 - \Phi(z_\alpha - c\psi_1(2))$.

From Proposition 2.3.3, we notice that the asymptotic power of both Fisher's method and Edgington's method depends on the constant c. Figure 2.4 shows the asymptotic power of the two methods at the significance level $\alpha = 0.05$. The figure indicates that Edgington's method has better power than Fisher's method under all considered constants. The difference between the power is maximized when the constant c is 2.81 and the difference is less than 0.001 when the constant c is less than 0.04 or greater than 7.35. We also notice that both methods have asymptotic power 1 as c increases.

Remark 2.3.1 *Proposition 2.3.3 implies that Edgington's method is more asymptotically powerful than Fisher's method if* $z_\alpha - \dfrac{c}{4\sqrt{1/12}}$ *is less than* $z_\alpha - c\psi_1(2)$. *Hence, under hypotheses (2.1), Edgington's method has better power than Fisher's method regardless of the value c because* $\dfrac{1}{4\sqrt{1/12}}$ *is greater than* $\psi_1(2)$.

Remark 2.3.2 *If considering the alternative hypothesis* $H_A : \theta = 1 + cp^{-a}$, *we cannot asymptotically detect departures from the null when* $a > 1/2$ *and would have asymptotic power 1 when* $0 < a < 1/2$ *and c is positive.*

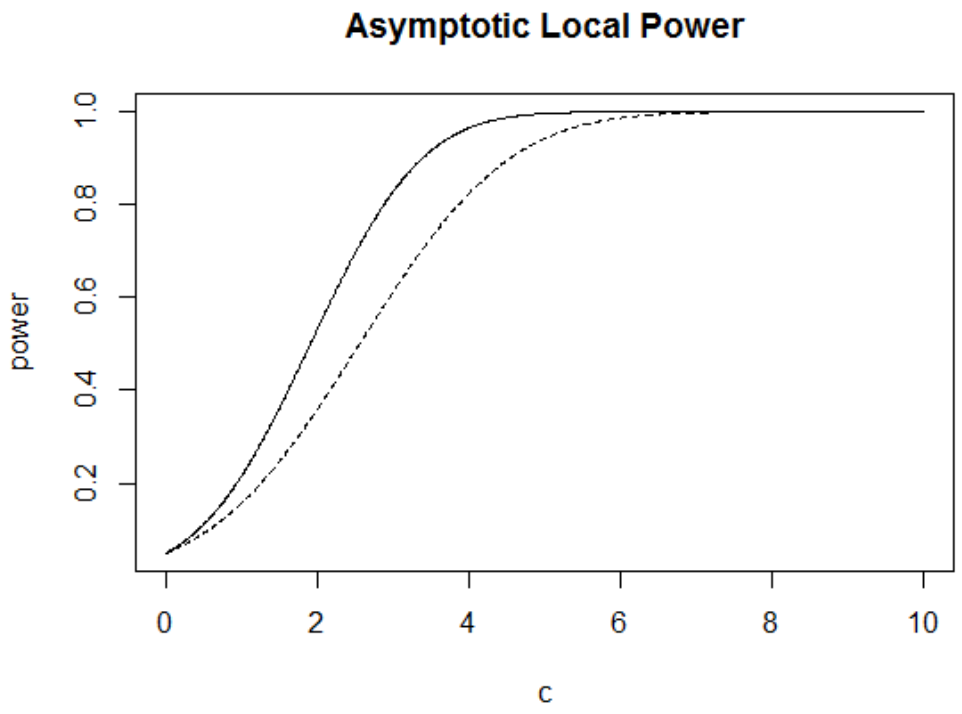

Figure 2.4: This figure shows the asymptotic power of Edgington's method and Fisher's method under hypotheses (2.1). The solid and dashed lines represent Edgington's method and Fisher's method, respectively.

Proposition 2.3.4 *The asymptotic powers of Edgington's method and Fisher's method under hypotheses (2.2) are* $\Phi\left(-z_\alpha - \frac{E_g(P) - 1/2}{\sqrt{1/12}}\right)$ *and* $1-\Phi\left(z_\alpha - \frac{E_g(-2\log P) - 2}{\sqrt{4}}\right)$ *, respectively, where* E_g *denotes expectation under the density function* g*.*

Proof By following the proof of Proposition 2.3.3, for Edgington's method, we just need to consider $E_{H_A}(P)$ and $\mathrm{Var}_{\mathrm{H_A}}(\mathrm{P})$. Under the considered alternatives, $E_{H_A}(P) = E_{H_0}(P) + p^{-1/2}(E_g(P) - E_{H_0}(P))$ and $\mathrm{Var}_{H_A}(P) = \mathrm{Var}_{H_0}(P) + o_p(1)$. By using these facts,

$$\frac{\sqrt{1/12}}{\sqrt{\mathrm{Var}_{H_A}(P)}}\left(-z_\alpha - \frac{\sqrt{p}(E_{H_A}(P) - 1/2)}{\sqrt{1/12}}\right) = (1 + o_p(1))\left(-z_\alpha - \frac{(E_g(P) - 1/2)}{\sqrt{1/12}}\right)$$

Hence, asymptotic power of Edgington's method under the considered alternatives is equal to $\Phi\left(-z_\alpha - \dfrac{E_g(P) - 1/2}{\sqrt{1/12}}\right)$. Similarly, for Fisher's method, $E_{H_A}(-2\log P)$ and $\mathrm{Var}_{H_A}(-2\log P)$ are necessary to be computed.

Under the considered alternatives,

$E_{H_A}(-2\log P) = E_{H_0}(-2\log P) + p^{-1/2}(E_g(-2\log P) - E_{H_0}(-2\log P))$ and

$\mathrm{Var}_{H_A}(P) = \mathrm{Var}_{H_0}(-2\log P) + o_p(1)$.

By using these,

$$\begin{aligned}&\frac{\sqrt{4}}{\sqrt{\mathrm{Var}_{H_A}(-2\log P)}}\left(z_\alpha - \frac{\sqrt{p}(E_{H_A}(-2\log P) - 2)}{\sqrt{4}}\right)\\ &= (1 + o_p(1))\left(z_\alpha - \frac{(E_g(-2\log P) - 2)}{\sqrt{4}}\right).\end{aligned}$$

Hence, the asymptotic power of Fisher's method is $1 - \Phi\left(z_\alpha - \dfrac{(E_g(-2\log P) - 2)}{\sqrt{4}}\right)$.

Remark 2.3.3 *For Edgington's method, the asymptotic power under hypotheses (2.2) increases as* $\dfrac{E_g(P) - 1/2}{\sqrt{1/12}}$ *decreases. For Fisher's method, on the contrary, the asymptotic power increases as* $\dfrac{E_g(-2\log P) - 2}{\sqrt{4}}$ *increases.*

Remark 2.3.4 *Proposition 2.3.4 indicates that Edgington's method is more powerful than Fisher's method when* $1 - 2E_g(P)$ *is greater than* $\sqrt{1/12}(E_g(-2\log P) - 2)$ *when the one-sided test is used. For future reference, we will call* $1 - 2E_g(P)$ *and* $\sqrt{1/12}(E_g(-2\log P) - 2)$ *mean differences.*

Remark 2.3.5 *From Propositions 2.3.3 and 2.3.4, we notice that the power of both Fisher's method and Edgington's method depends on the first two moments of P-values. Hence, we will call these methods "moment based tests".*

Remark 2.3.6 *Similarly to Remark 2.3.2, we notice that if we have an alternative hypothesis,* $H_A : f = (1 - p^{-a})I_{(0,1)}(P) + p^{-a}g$*, moment based tests cannot asymptot-*

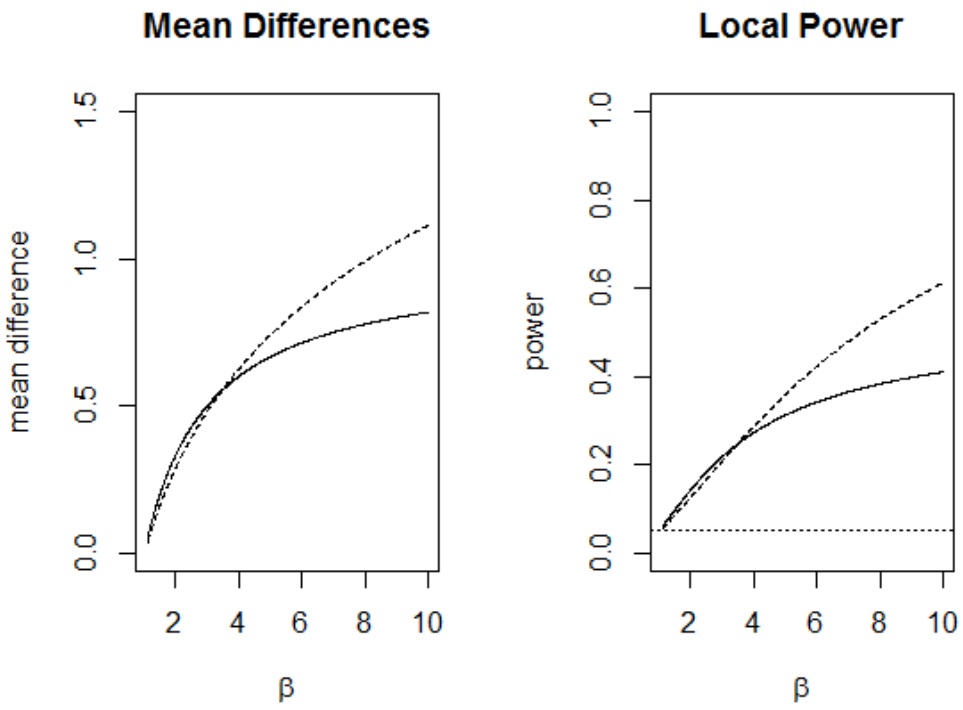

Figure 2.5: This figure shows mean differences and asymptotic power under $\sqrt{p}$-alternatives in hypotheses (2.2) when g is beta distributions with parameters $\alpha = 1$ and $\beta > 1$. The solid and dashed lines represent the power of Edgington's method and Fisher's method, respectively. The horizontal dotted line in the right plot represents the level of tests 0.05.

ically detect departures from the null when $0 < a < 1/2$ *and would have asymptotic power 1 when* $a > 1/2$ *and there is no bias problem.*

Remark 2.3.7 *Even if the bias problem exists, a moment based test has asymptotic power 1 when* $a > 1/2$ *as long as the two-sided test is used.*

To compare the power of the two methods under hypotheses (2.2), beta distributions with parameters $\alpha = 1$ and $\beta > 1$ are considered. Note that we do not have the bias problem by considering such beta distributions, and we will not consider the two-sided test at this point. Figure 2.5 shows the mean differences and the power

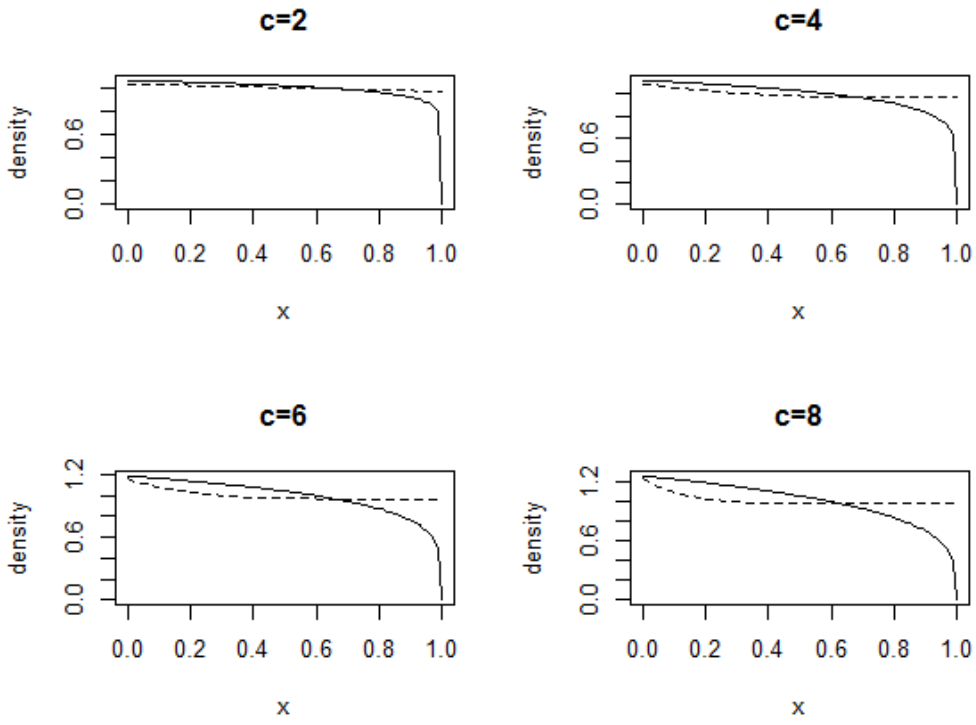

Figure 2.6: This figure shows the density of the two alternative hypotheses when there are 1,000 data sets. The solid line represents the density of alternative distribution under hypotheses (2.1). The dashed line represents the mixture of uniform and beta(1,c), which corresponds to alternative distribution under hypotheses (2.2).

of the two methods for various parameters β from 1.1 to 10 at the significance level α=0.05. In the figure, the solid and dashed lines denote the power of Edgington's method and Fisher's method, respectively. When β is less than 3.45, Edgington's method is slightly better than Fisher's method. Fisher's method has higher power than Edgington's method when β is greater than 3.45 and the difference in the power gets bigger as β increases. Such a phenomenon suggests that Fisher's method might be preferable to Edgington's method when there exists strong evidence against the null. However, when we have data sets with small sample sizes, it may be difficult to expect that we have strong evidence against the null. Hence, this result may

imply that Edgington's method is preferable to Fisher's method in the setting of this dissertation. This result agrees with the Pitman efficiency result from Proposition 2.3.2.

Under hypotheses (2.1), both Edgington's method and Fisher's method have asymptotic power 1 as the constant c increases. However, under hypotheses (2.2), the two methods do not have asymptotic power 1 even if the evidence against the null gets stronger. Such a difference in behavior of the two methods might be explained by the shape of the two alternative distributions. Figure 2.6 shows the density of alternative distributions. From the figure, we notice that the alternative distribution under hypotheses (2.1) shows more deviations from uniformity than those under hypotheses (2.2). Especially, the density of alternative distributions from hypotheses (2.1) tends to approach 0 for large P-values, indicating a severe departure from uniformity. This may be the reason that the asymptotic power of both methods under hypotheses (2.1) is 1 as c increases, unlike the asymptotic power under hypotheses (2.2).

It is clear that power would decrease if we apply the two-sided tests when tests are not biased. The relative power decrease, defined as the ratio of the power of two-sided tests subtracted from that of the one-sided test to the power of the one-sided test, are investigated. Hence, negative values of the relative power decrease imply that power increases as the result of applying the two-sided test. Figures 2.7 and 2.8 show the relative power decrease at various significance levels α_2 under hypotheses (2.1) and (2.2), respectively. Under hypotheses (2.1), Fisher's method tends to lose more power when the constant is greater than 1 at all considered significance levels α_2. Under hypotheses (2.2), as β and the significance level α_2 increases, the amount of relative power decrease tends to get larger. Especially, the decrease in power is at least 2%, 4%, 6% and 9% when α_2 is 0.005, 0.01, 0.015, and 0.02, respectively.

On the contrary, power would increase if we apply the two-sided tests when tests are biased. To consider the bias of test, for hypotheses (2.1), negative constants c are used and, for hypotheses (2.2), beta distributions with parameters $\alpha > 1$ and $\beta = 1$ are used as the density g. Figures 2.9 and 2.10 show power of the two methods under hypotheses (2.1) and (2.2), respectively. We notice that the power of both methods are less than the size of the test 0.05 when the one-sided test is used. We also notice that there exists a serious bias problem as parameter α increases or the constant c decreases. To investigate the effect of the two-sided tests, power of the two-sided tests at significance level α_2 is obtained. Figures 2.11 and 2.12 show the asymptotic power of the two-sided tests. Under both considered alternatives, when there exists a serious bias problem, we notice that the bias is corrected at a small α_2, such as 0.005. Especially, under hypotheses (2.1), the power is close to 1 regardless the significance level α_2. However, under both hypotheses, the bias problem is not resolved at relatively large α_2, such as 0.15 and 0.20 when there is a mild bias problem. We notice that Fisher's method tends to have lower power than Edgington's method when the two-sided test is applied. This might suggest that Edgington's method is better when tests are biased and the two-sided test is used.

We investigate the effects of the two-sided tests from Figures 2.7 to 2.12. This suggests that we need to consider both possible decrease in power and the chance of the bias correction which result from applying two-sided tests. Clearly, these results depend on the significance level α_2, and a cautious choice of the significance levels is essential. Even if there is no obvious solution regarding the choice of the significance levels, it might be safe to use α_2 less than or equal to 0.015 by considering the effects of two-sided tests.

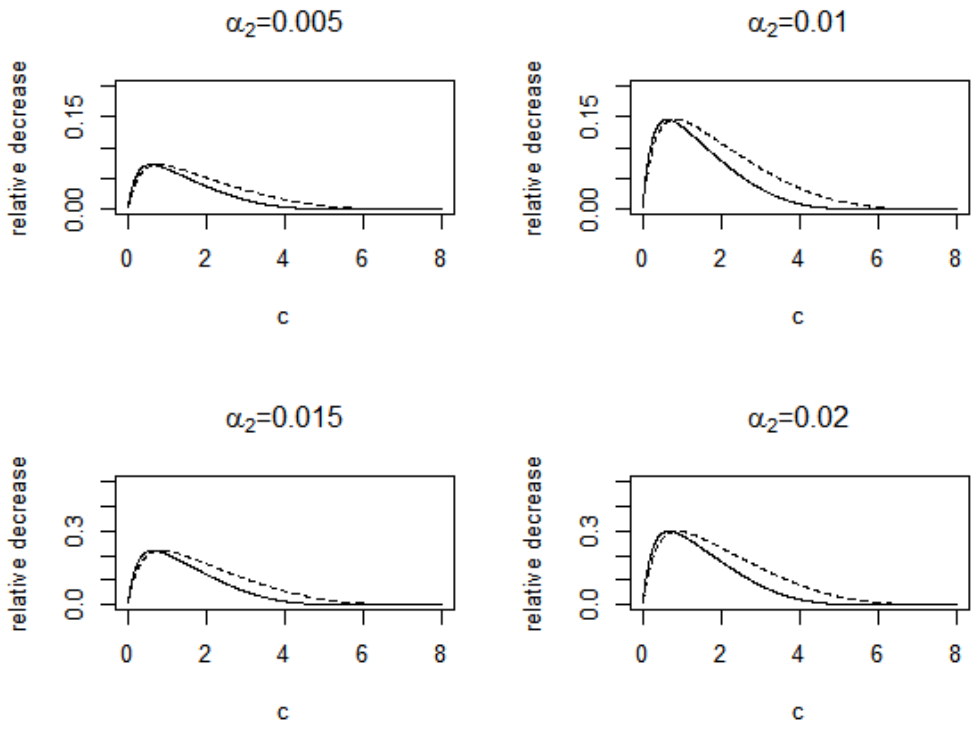

Figure 2.7: This figure shows the relative decrease in the asymptotic power under hypotheses (2.1) when tests are not biased and the two-sided test is used. The solid and dashed lines represent the relative power decrease of Edgington's method and Fisher's method, respectively.

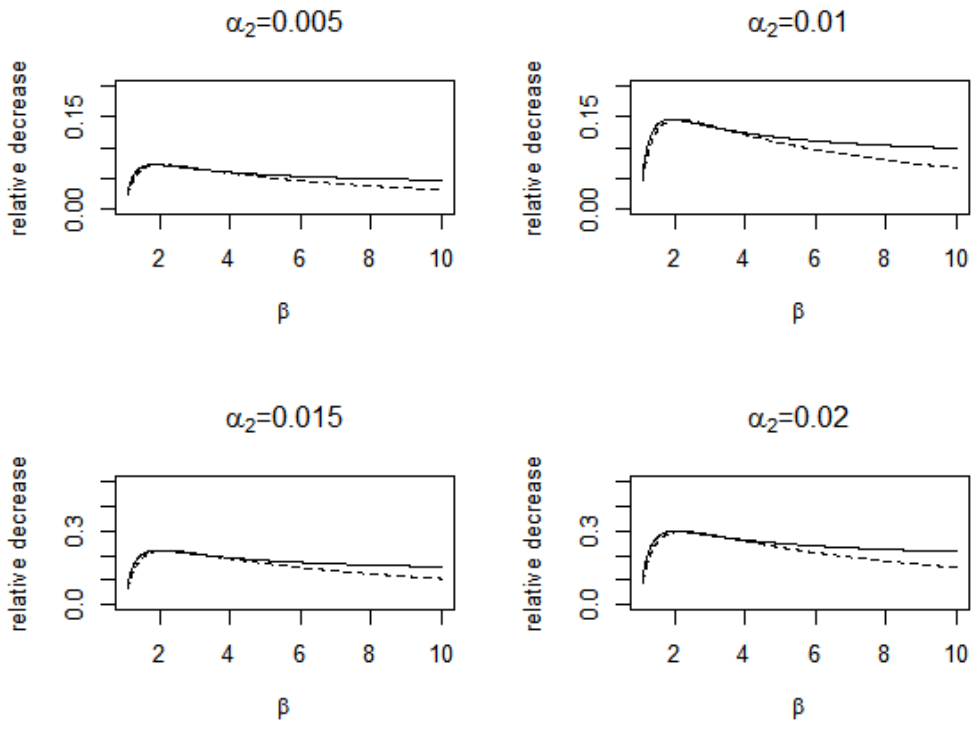

Figure 2.8: This figure shows the relative decrease in the asymptotic power under hypotheses (2.2) when tests are not biased and the two-sided test is used. The solid and dashed lines represent the relative power decrease of Edgington's method and Fisher's method, respectively.

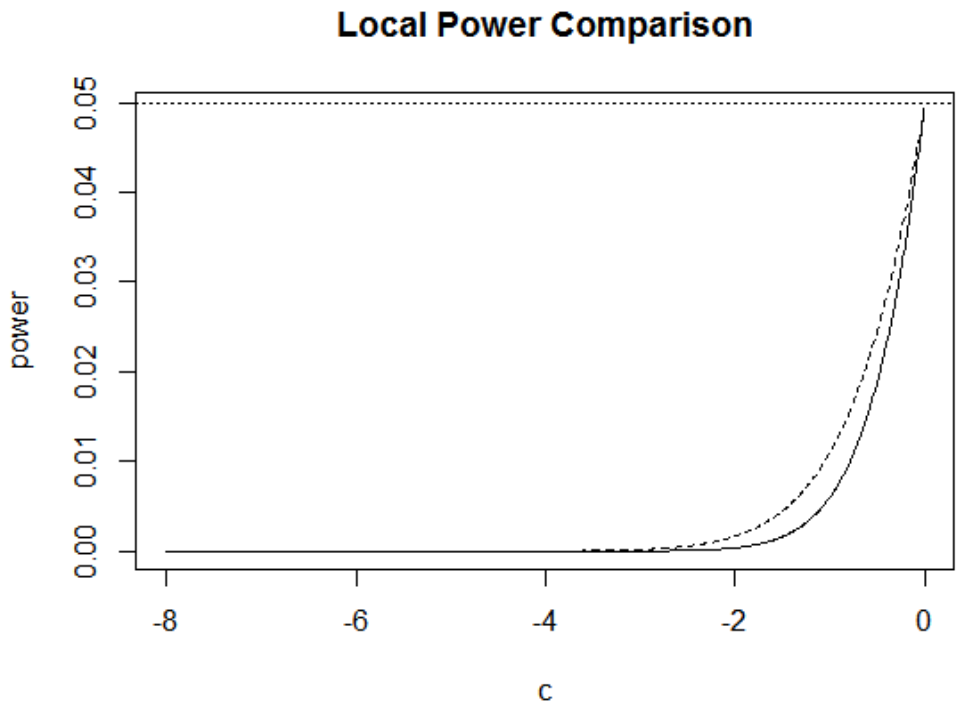

Figure 2.9: This figure shows the asymptotic power under hypotheses (2.1) when tests are biased and one-sided test is used. The solid and dashed lines represent the power of Edgington's method and Fisher's method, respectively.

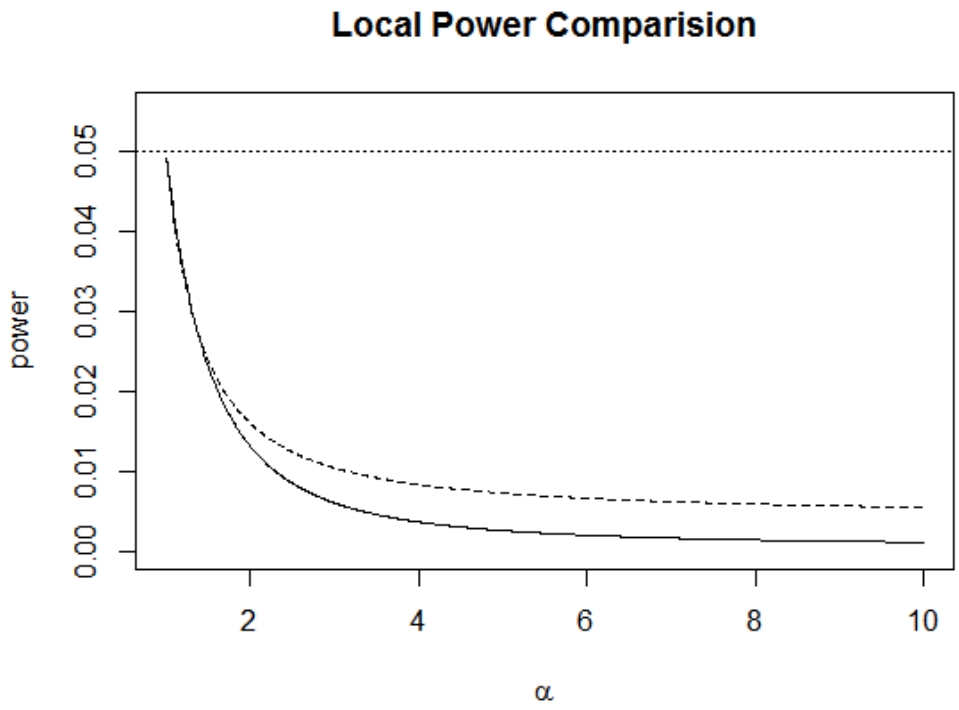

Figure 2.10: This figure shows the asymptotic power under hypotheses (2.2) when tests are biased and one-sided test is used. The solid and dashed lines represent the power of Edgington's method and Fisher's method, respectively.

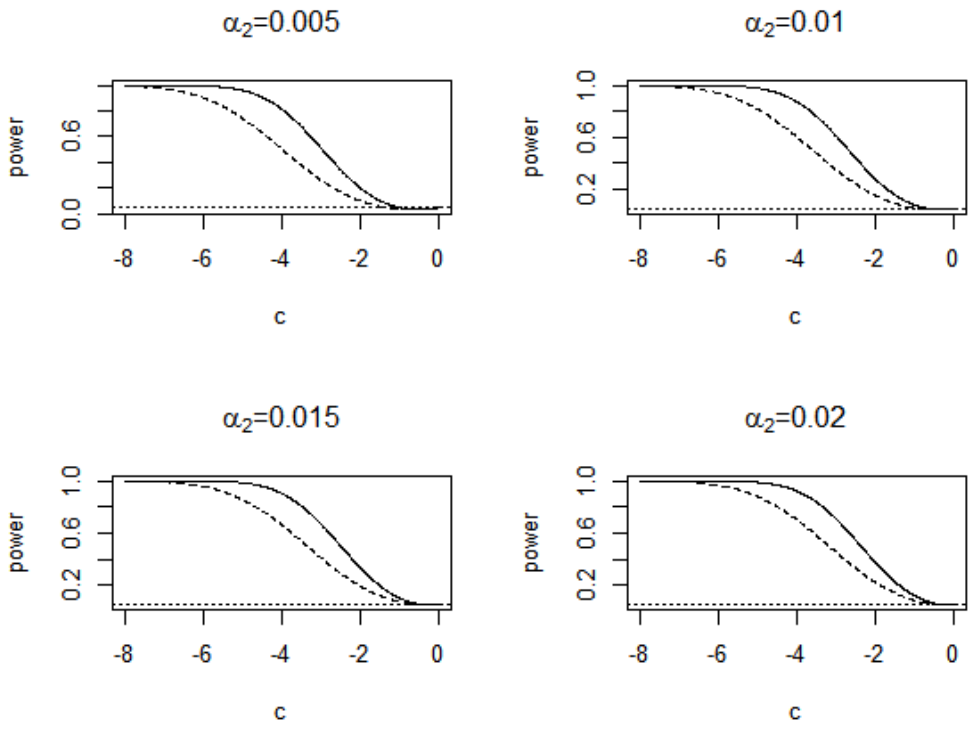

Figure 2.11: This figure shows asymptotic power of the two-sided tests under hypotheses (2.1) when tests are biased. The solid and dashed lines represent the power of Edgington's method and Fisher's method, respectively. The dotted line denotes the significance level $\alpha = 0.05$.

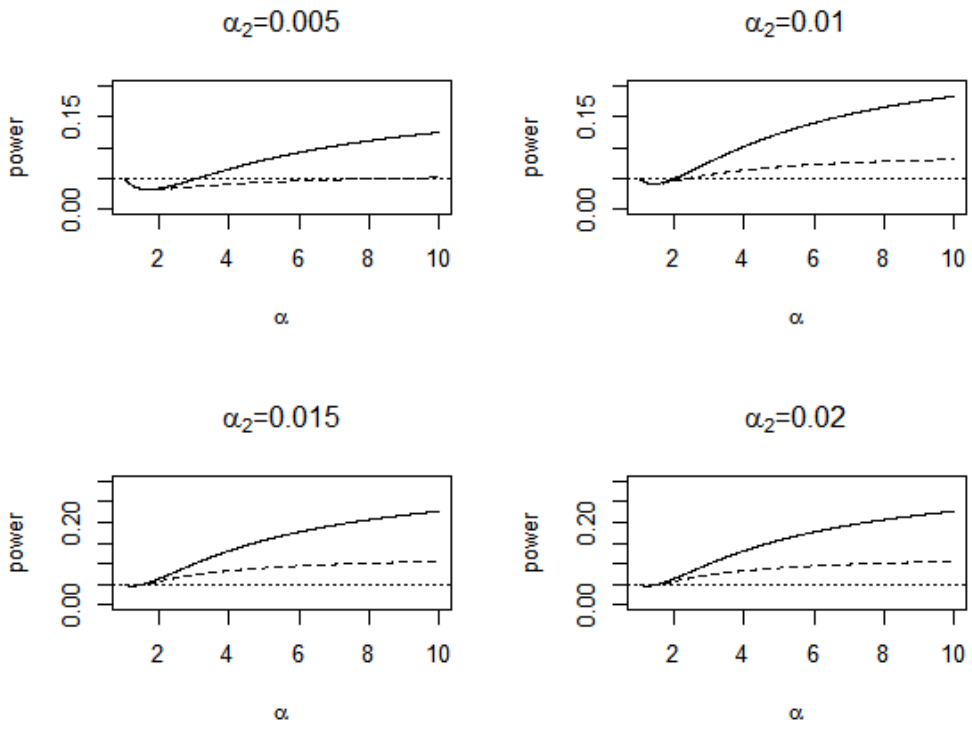

Figure 2.12: This figure shows asymptotic power of the two-sided tests under hypotheses (2.2) when tests are biased. The solid and dashed lines represent the power of Edgington's method and Fisher's method, respectively. The dotted line denotes the significance level $\alpha = 0.05$.

2.4 Other methods of combining test results

In addition to methods of combining P-values using Fisher's method or Edgington's method, other possible ways to combine results from small data sets is to apply the order selection test (Hart, 1997; Kim, 2000) or the smooth test (Kallenberg and Ledwina, 1997; Ledwina, 1994; Inglot and Ledwina, 2006) by using the fact that the distribution of the P-value under the null hypothesis follows the uniform distribution.

The smooth test (Neyman, 1937) postulates an alternative hypothesis that includes the uniform distribution as a special case. An order k alternative probability density function $g_k(x)$ is defined by

$$g_k(x) = C(\theta) \exp\left\{\sum_{i=1}^{k} \theta_i h_i(x)\right\}$$

where $C(\theta)$ is a normalizing constant and $\{h_i(x)\}$ is a set of orthonormal functions.

When considering an order k alternative hypothesis, the smooth test is equivalent to testing the null hypothesis, $H_0 : \theta_1 = \cdots = \theta_k = 0$. Neyman recommended using a score test statistic, $S_k = \sum_{i=1}^{k} U_i^2$ where $U_i = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} h_i(X_j)$. He also suggested that four components would be enough. Kallenberg and Ledwina (1997) suggested a data-driven smooth test which selects the order using BIC. Hence, the data-driven smooth test is composed of two parts; one is a selection rule to choose an appropriate sub-model and the other is the score test statistic corresponding to the selected order. The performance of the test depends on characteristics of the selection rule. When BIC is used, the test will have poor power for "high frequency" alternatives, i.e., alternatives for which all θ_i are 0 except those at large values of i. This is due to the relatively large penalty that BIC imposes on models with k large. If AIC is used instead of BIC, the test is better at detecting high frequency alternatives since AIC

penalizes models less severely than does BIC. To compromise between these two selection rules, Inglot and Ledwina (2006) suggested a test which uses AIC when the distribution is far from the null and BIC when the distribution is not far from the null. Specifically, they considered the distribution to be far from the null when $\max_{1\leq i\leq d(p)} |U_i|$ is greater than $\sqrt{2\log p}$ where p and $d(p)$ denote the sample size and the maximum of considered orders, respectively. Their simulation shows that the power of their test is between the power of the test based on AIC and that of the test based on BIC. For example, if an alternative distribution is close to the null distribution, the test based on AIC tends to have lower power than the test based on BIC. The suggested test has power between those two powers, indicating that this test cannot have the best power. On the other hand, the test depending on the modified selection rule has an advantage over the test which merely uses AIC or BIC since it has better power than the worst of the AIC and BIC based tests.

The order selection test postulates an alternative distribution represented by the Fourier series

$$g(x) = 1 + 2\sum_{i=1}^{\infty} \phi_j \cos(\pi j x), \text{ where } \phi_j = \int_0^1 g(x)\cos(\pi j x)\mathrm{d}x, j = 1, 2, \ldots$$

The Fourier coefficient ϕ_j may be estimated by $\hat{\phi}_j = \frac{1}{p}\sum_{i=1}^{p}\cos(\phi_j P_i), j = 1, 2, \ldots$ Defining S_p by

$$S_p = \max_{1\leq m\leq p} \frac{1}{m}\sum_{j=1}^{m} 2p\hat{\phi}_j^2,$$

the order selection test rejects the null for large values of S_p. Kim (2000) shows that, under the null, the test statistic has the limiting distribution

$$F_{OS}(\gamma) = \exp\left\{-\sum_{j=1}^{\infty}\frac{P(\chi_j^2 > j\gamma)}{j}\right\},$$

where χ^2_j is a chi-squared random variable with j degrees of freedom. He also found γ which guarantees the right size of test by simulations. At the significance level 0.05, γ is 4.18. Hence, the suggested test is asymptotically equivalent to rejecting the null hypothesis when $S_n \geq 4.18$.

Both order selection test and data-driven smooth test detect smooth departures from the null. Hence, we will call these tests smoothing based tests. Even if both tests detect smooth departures from the null, each one has its advantages. For example, the smooth test may be more powerful than the order selection test when the alternative distribution deviates little from the null because the score test is the most powerful test for small deviations. Such a property of score tests follows from the fact that, when h is small, $L(\theta_0 + h) \simeq L(\theta_0) + hL'(\theta_0)$, where L is the log-likelihood function and θ_0 is the true parameter value under the null. On the contrary, the order selection test tends to have higher power than the smooth test when the alternative distribution is of high frequency type.

The smoothing based tests have an advantages over moment based tests in the sense that they detect any sort of departure of the P-value distribution from uniformity. For example, assume that the distribution of the P-value is beta(1.11,1.11). The density in Figure 2.13 shows only a small departure from the null. Under this distribution for the P-value, Edgington's method cannot consistently detect the alternative because the mean of the beta distribution is the same as that of the uniform distribution. If the two-sided test is used, Fisher's method will have better power than Edgington's method under the alternative distribution since the mean of $-2 \log P$ under the beta distribution is 1.936, which is different from 2. However, the method still does not have good power. For example, when there are 500 small data sets and a two-sided test is applied by using two significance levels, $\alpha_1 = 0.04$ and $\alpha_2 = 0.01$, the power of Fisher's method is just 0.066. This indicates that even if

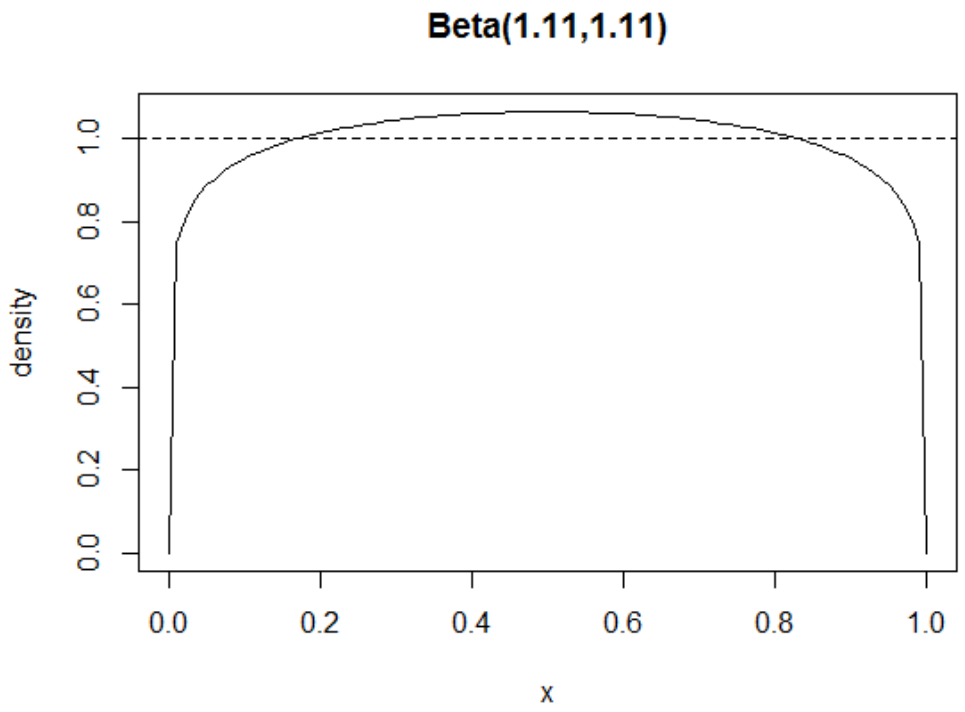

Figure 2.13: This figure shows the density of the beta(1.11,1.11).

there exists a departure from the null, moment based tests do not have good power against the null as long as the first two moments under the alternative are close to those under the null. In this case, smoothing based tests might have better power than moment based tests. However, moment based tests may have an advantage over smoothing based tests for selected alternatives, especially when the first two moments of the distribution of P-values are quite different from the moments of the uniform distribution. It is clear that both kinds of tests have their desirable properties, and hence, we will use both tests and compare the power instead of choosing one test.

2.5 Asymptotic distribution theory for moment based test

When sample size is small, the exact null distributions of edf-based gof test statistics are unknown. To obtain the P-value, the null distribution of the test statistic can be approximated via simulation. It is well-known that the edf converges uniformly to the true distribution function with probability 1 by the Glivenko-Cantelli theorem. However, this theorem cannot guarantee that the asymptotic null distribution based on empirical P-values is equivalent to that based on the true P-value. In this section, we will show that these two asymptotic null distributions are the same. Also, the effect of the number of bootstrap replications and data sets will be discussed for Edgington's method.

We assume that $T_1, \ldots, T_p$ and $Y_1, \ldots, Y_N$ are test statistics from small data sets and from simulations, respectively. Under the null, both are independent and identically distributed as F, where F is the null distribution of the test statistic.

Theorem 2.5.1 *Under the null, $S_{N,p}$ and $U_p + V_N - \frac{1}{2}$ have the same asymptotic distribution as $N \to \infty$ and $p \to \infty$, where $S_{N,p} = \frac{1}{p}\sum_{i=1}^{p}(1 - \hat{F}_N(T_i))$, $\hat{F}_N(x) = \frac{1}{N}\sum_{j=1}^{N} I(Y_j \le x)$, $U_p = \frac{1}{p}\sum_{i=1}^{p}(1 - F(T_i))$ and $V_N = \frac{1}{N}\sum_{j=1}^{N} F(Y_j)$.*

Proof It is sufficient to show that $\dfrac{E\left[\left(S_{N,P} - U_p - V_N + \frac{1}{2}\right)^2\right]}{\mathrm{Var}(S_{N,p})}$ converges to 0 as $N \to \infty$ and $p \to \infty$.

$$
\begin{aligned}
&E\left[\left(S_{N,p} - U_p - V_N + \frac{1}{2}\right)^2\right] \\
&= E[(S_{N,p} - U_p)^2] - 2E[(S_{N,p} - U_p)(V_n - \frac{1}{2})] + E[(V_N - \frac{1}{2})^2] \\
&= A_1 - 2A_2 + A_3
\end{aligned}
$$

where $A1 := E[(S_{N,p} - U_p)^2], A_2 := E[(S_{N,p} - U_p)(V_N - \frac{1}{2})]$ and $A_3 := E[(V_N - \frac{1}{2})^2]$.

By the law of total expectation,

$$\begin{aligned}
A_1 &= E[(S_{N,p} - U_p)^2] \\
&= E\left[E\left[\left(\frac{1}{p}\sum_{i=1}^{p}\hat{F}_N(T_i) - \frac{1}{p}\sum_{i=1}^{p}F(T_i)\right)^2 \middle| T_1, \ldots, T_p\right]\right] \\
&= E\left[\mathrm{Var}\left(\frac{1}{p}\sum_{i=1}^{p}\hat{F}_N(T_i)\middle| T_1, \ldots, T_p\right)\right] \\
&= E\left[\frac{F(T_1)(1-F(T_1))}{Np} + \frac{p-1}{p}\frac{F(\min(T_1,T_2)) - F(T_1)F(T_2)}{N}\right] \\
&= \frac{1}{12N} + \frac{1}{12Np}.
\end{aligned}$$

The last equation can be obtained by direct calculation of expectation under the null and by using the fact that the density of the minimum of two uniform random variables is $2(1-u)$, where $0 < u < 1$. Therefore, $E[F(T_1)(1-F(T_1))] = \frac{1}{6}$ and $E[F(\min(T_1,T_2))] = \frac{1}{3}$.

Similarly, we can also use the law of total expectation to obtain A_2, which is

$$\begin{aligned}
A_2 &= E\left[E\left[(S_{N,p} - U_p)(V_N - \frac{1}{2}) \mid Y_1, \ldots, Y_N\right]\right] \\
&= E\left[(V_N - \frac{1}{2})\{E(S_{N,p} \mid Y_1, \ldots, Y_N) - \frac{1}{2}\}\right] \\
&= E\left[(V_N - \frac{1}{2})(V_N - \frac{1}{2})\right] \\
&= E\left[(V_N - \frac{1}{2})^2\right] \\
&= \frac{1}{12N}.
\end{aligned}$$

Immediately above we have used

$$E(S_{N,p}|Y_1 \dots Y_N) = \frac{1}{N}\sum_{j=1}^{N} E\left(\frac{1}{p}\sum_{i=1}^{p} I(Y_j > T_i))|Y_1, \dots, Y_N\right)$$
$$= \frac{1}{N}\sum_{j=1}^{N} F(Y_j).$$

Since $A_3 = \dfrac{1}{12N}$, we obtain that $E\left[\left(S_{N,p} - U_p - V_N + \dfrac{1}{2}\right)^2\right] = \dfrac{1}{12Np}$.
Now, we need to find the order of $\mathrm{Var}(S_{N,p})$, which is

$$\mathrm{Var}(S_{N,p}) = E[\mathrm{Var}(S_{N,p}|T_1, \dots, T_p)] + \mathrm{Var}[E(S_{N,P}|T_1, \dots, T_p)]$$
$$= \frac{1}{12Np} + \frac{1}{12N} + \frac{1}{12p}.$$

$E[\mathrm{Var}(S_{N,p}|T_1, \dots, T_p)]$ is obtained from A_1 and $\mathrm{Var}[E(S_{N,p}|T_1, \dots, T_p)]$ can be computed easily by using $E(S_{N,P}|T_1, \dots, T_p) = \dfrac{1}{p}\displaystyle\sum_{i=1}^{p}(1 - F(T_i))$. Hence,
$E\left[\left(S_{N,P} - U_p - V_N + \dfrac{1}{2}\right)^2\right]$ is of smaller order than $\mathrm{Var}(S_{N,p})$.

The next corollary shows that the asymptotic null distribution of Edgington's statistic using empirical P-values is the same as that using the true P-values when $N \to \infty$, $p \to \infty$ and $p = o(N)$.

Corollary 2.5.2 *Under the null, if* $\dfrac{p}{N} \to c$, *where* $c > 0$ *is a constant when* $N \to \infty$ *and* $p \to \infty$ *then* $\dfrac{\sqrt{p}\left(\sum_{i=1}^{p}(1 - \hat{F}_N(T_i)) - \frac{1}{2}\right)}{\sqrt{1/12}}$ *converges to a normal distribution with mean 0 and variance* $1 + c$ *as* $N \to \infty$ *and* $p \to \infty$.

Proof Let $\sigma_{N,p}$ be the standard deviation of $S_{N,p}$. Then, $\dfrac{S_{N,p} - \frac{1}{2}}{\sigma_{N,p}}$ and $\dfrac{U_p - \frac{1}{2}}{\sigma_{N,p}} + \dfrac{V_N - \frac{1}{2}}{\sigma_{N,p}}$ have the same asymptotic distribution by Theorem 2.5.1. We know that

$\sigma_{N,p} = \sqrt{\frac{1}{12Np} + \frac{1}{12N} + \frac{1}{12p}}$. By using this, we can show that $\frac{\sqrt{1/12p}}{\sigma_{N,p}}$ and $\frac{\sqrt{1/12N}}{\sigma_{N,p}}$ converge to $\frac{1}{\sqrt{1+c}}$ and $\sqrt{\frac{c}{1+c}}$, respectively. Since $\frac{U_p - \frac{1}{2}}{\sigma_{N,p}}$ can be written as $\frac{\sqrt{p}(U_p - \frac{1}{2})}{\sqrt{1/12}} \frac{\sqrt{1/12p}}{\sigma_{N,p}}$, Slutsky's theorem implies that $\frac{U_p - \frac{1}{2}}{\sigma_{N,p}}$ converges to a normal distribution with mean 0 and variance $\frac{1}{1+c}$. Similarly, $\frac{V_N - \frac{1}{2}}{\sigma_{N,p}}$ converges to a normal distribution with mean 0 and variance $\frac{c}{1+c}$. Since U_p and V_N are independent, $\frac{S_{N,p} - \frac{1}{2}}{\sigma_{N,p}}$ converges to the standard normal distribution. The corollary now follows from Slutsky's theorem.

Corollary 2.5.2 shows the necessity of adjusting a critical value when the number of bootstrap replications is not large enough relative to the number of data sets. For example, when we have 1,000 data sets and 2,000 bootstrap replications, Corollary 2.5.2 suggests that the actual level of a nominal 0.05 test is 0.09. If the critical value is not adjusted, the test tends to reject the null more frequently than its predetermined significance level. The next theorem provides conditions under which Fisher's method based on the empirical P-values has the same asymptotic null distribution as that based on the true P-values.

Theorem 2.5.3 *If* $\frac{p}{\sqrt{N}}$ *converges to 0 as* N *and* p *tend to* ∞, $-2\sum_{i=1}^{p} \log(1 - \hat{F}(T_i))$ *and* $-2\sum_{i=1}^{p} \log(1 - F(T_i))$ *have the same asymptotic null distribution.*

Proof Let $\hat{P}_i$ be $1 - \hat{F}(T_i)$ and P_i be $1 - F(T_i)$, where $\hat{F}$ and F are the empirical and true distribution functions of test statistics under the null. By Taylor's expansion, $\log \hat{P}_i = \log P_i + \tilde{P}_i^{-1}(\hat{P}_i - P_i)$, where $\tilde{P}_i$ is between P_i and $\hat{P}_i$.
We have

$$\left|\sum_{i=1}^{p} \log \frac{\hat{P}_i}{P_i}\right| = \left|\sum_{i=1}^{p} \tilde{P}_i^{-1}(\hat{P}_i - P_i)\right| \leq (\min \tilde{P}_i)^{-1} \sum_{i=1}^{p} |\hat{P}_i - P_i| \leq \frac{p}{\tilde{P}_{(1)}} \sup_i |\hat{P}_i - P_i|,$$

where $\tilde{P}_{(1)} = \min_i \tilde{P}_i$. Let r be such that $\tilde{P}_{(1)}$ is between $\hat{P}_r$ and P_r. There are two cases; one is $\hat{P}_r < \tilde{P}_{(1)} < P_r$ and the other is $P_r < \tilde{P}_{(1)} < \hat{P}_r$.

For the first case,

$$\tilde{P}_{(1)} = P_r + \tilde{P}_{(1)} - P_r \geq P_{(1)} - \sup_i |\tilde{P}_i - P_i| = P_{(1)} - M_p,$$

where $P_{(1)} = \min_i P_i$ and $M_p = \sup_i |\tilde{P}_i - P_i|$. This implies that

$$\left| \sum_{i=1}^{p} \log \frac{\hat{P}_i}{P_i} \right| \leq \frac{p}{P_{(1)} - M_p} \sup_i |\hat{P}_i - P_i|.$$

For the second case, $\tilde{P}_{(1)} > P_{(1)}$. For this case, $\left| \sum_{i=1}^{p} \log \frac{\hat{P}_i}{P_i} \right| \leq \frac{p}{P_{(1)}} \sup_i |\hat{P}_i - P_i|$. Hence, we need to choose N so large that

(i) $P_{(1)} - M_p$ is asymptotic to $P_{(1)}$, and

(ii) $\frac{p}{P_{(1)}} \sup_i |\hat{P}_i - P_i|$ is of smaller order than p, which is the order of $-2\sum_{i=1}^{p} \log P_i$.

To show (i), the order of $P_{(1)}$ is obtained first. By noting that $P(pP_{(1)} < x) = 1 - P(P_{(1)} > x/p) = 1 - (1 - x/p)^p$, it is easily shown that the order of $P_{(1)}$ is $\frac{1}{p}$ because $P(pP(1) < x)$ converges to an exponential random variable with rate $\alpha = 1$. Since $\sup_i |\tilde{P}_i - P_i| \leq \sup_i |\hat{P}_i - P_i|$ and the order of $\sup_i |\hat{P}_i - P_i|$ is $\frac{1}{\sqrt{N}}$ by Donsker's theorem, we see that $P_{(1)} - \sup_i |\tilde{P}_i - P_i|$ is asymptotic in probability to $P_{(1)}$ if $\frac{p}{\sqrt{N}}$ converges to 0 as N and p increase without bound. To show (ii), it is enough to verify that $\frac{1}{P_{(1)}} \sup_i |\hat{P}_i - P_i|$ converges to 0 under the given condition, and this holds by the Glivenko-Cantelli theorem. This implies that $-2\sum_{i=1}^{p} \log \hat{P}_i$ and $-2\sum_{i=1}^{p} \log P_i$ have the same asymptotic distribution under the null as long as the given condition is satisfied.

The condition in Theorem 2.5.3 is strong, especially when we have a large number of data sets. For example, when we have 1,000 data sets, the number of bootstrap replications N should be of the order $10^6 \log 10^3$. Too many bootstrap replications require excessive computing time. However, while the condition is sufficient, it may not be necessary to obtain the result in Theorem 2.5.3. Through simulations in Chapter 3, we find that 100,000 is usually enough to obtain a good approximation of the null distribution when we have 1,000 data sets.

2.6 Asymptotic power for local alternatives

In this section, we will show that Edgington's method using the empirical P-values can detect $\sqrt{p}$-alternatives asymptotically. As in the previous section, $T_1, \ldots, T_p$ and $Y_1, \ldots, Y_N$ are test statistics from data sets and simulated statistics, respectively. We assume that, $T_1, \ldots, T_p$ are independent and identically distributed as

$$F_1(t) = (1 - \frac{1}{\sqrt{p}})F(t) + \frac{1}{\sqrt{p}}G(t), \tag{2.3}$$

where F is the null distribution of the test statistic and G is a distribution different than F. Of course, $Y_1, \ldots, Y_N$ are independent and identically distributed as F.

Theorem 2.6.1 *If each $T_i, i = 1, \ldots, p$ has distribution F_1 defined in (2.3) then the statistic $S_{N,p}$ and $U_p + V_N - \frac{1}{2}$ have the same asymptotic distribution as $N \to \infty$ and $p \to \infty$, where $S_{N,p}, U_p$ and V_N are defined in Theorem 2.5.1.*

Proof As in Theorem 2.5.1, it is sufficient to show that $\dfrac{E\left[\left(S_{N,P} - U_p - V_N + \frac{1}{2}\right)^2\right]}{\mathrm{Var}(S_{N,p})}$ converges to 0 as $N \to \infty$ and $p \to \infty$. A_1, A_2 and A_3 are defined in the proof of Theorem 2.5.1.

By the law of total expectation,

$$\begin{aligned} A_1 &= E[(S_{N,p} - U_p)^2] \\ &= E\left[\frac{F(T_1)(1-F(T_1))}{Np} + \frac{p-1}{p}\frac{F(\min(T_1,T_2)) - F(T_1)F(T_2)}{N}\right] \end{aligned}$$

Expectations can be found by direct computations:

$$E[F(T_1)(1-F(T_1)] = \frac{1}{\sqrt{p}}\int F(t)(1-F(t))dG(t) + \frac{1}{6}(1-\frac{1}{\sqrt{p}}),$$

$$\begin{aligned} E[F(T)] &= \int F(t)dG(t) \\ &= \frac{1}{\sqrt{p}}\int F(t)dG(t) + \left(1-\frac{1}{\sqrt{p}}\right)\frac{1}{2} \end{aligned}$$

and

$$\begin{aligned} E[F(\min(T_1,T_2))] &= 2\int F(u)(1-F_1(u))dF_1(u) \\ &= \frac{2}{\sqrt{p}}\int F(u)dG(u) - \frac{2}{p}\int F(u)G(u)dG(u) \\ &\quad - \frac{2}{\sqrt{p}}\int F(u)G(u)dF(u) - \frac{2}{\sqrt{p}}(1-\frac{1}{\sqrt{p}})\int F(u)^2 dG(u) \\ &\quad + \frac{1}{3} + \frac{1}{3\sqrt{p}} - \frac{2}{3p}. \end{aligned}$$

Therefore, $A_1 = \dfrac{1}{12N} + O\left(\dfrac{1}{N\sqrt{p}}\right)$.

We have

$$\begin{aligned} A_2 &= E[(S_{N,p} - U_p)(V_N - \frac{1}{2})] \\ &= E[(V_N - \frac{1}{2})E[S_{N,p}|Y_1,\ldots,Y_N]] \end{aligned}$$

$$
\begin{aligned}
&= E\left[(V_N - \frac{1}{2})\frac{1}{N}\sum_{j=1}^{N} F_1(Y_j)\right] \\
&= \frac{1}{\sqrt{p}} E\left[\frac{1}{N}\sum_{j=1}^{N} F(Y_j)\frac{1}{N}\sum_{j=1}^{N} G(Y_j)\right] + \left(1 - \frac{1}{\sqrt{p}}\right) E\left[\left(\frac{1}{N}\sum_{i=1}^{N}(\frac{1}{2} - F(Y_i))\right)^2\right] \\
&- \frac{1}{2\sqrt{p}} E[G(Y)] \\
&= \frac{1}{12N} + O\left(\frac{1}{N\sqrt{p}}\right).
\end{aligned}
$$

The second equation can be obtained by using $E[(V_N - \frac{1}{2})U_P] = 0$, and the third equation uses $E[S_{N,p}|Y_1, \ldots, Y_n] = \frac{1}{N}\sum_{j=1}^{N} E[I(Y_j > T_1)] = \frac{1}{N}\sum_{j=1}^{N} F_1(Y_j)$. Since A_3 is $\frac{1}{12N}$, $E\left[\left(S_{N,p} - U_p - V_N + \frac{1}{2}\right)^2\right]$ is of order $\frac{1}{N\sqrt{p}}$. Now, we need to find the order of $\mathrm{Var}(S_{N,p})$. To obtain it, we need to obtain $\mathrm{Var}[E(S_{N,p}|T_1, \ldots, T_p)]$, which is

$$
\begin{aligned}
\mathrm{Var}[E(S_{N,p}|T_1, \ldots, T_p)] &= \frac{1}{p}\mathrm{Var}(F(T_1)) \\
&= \frac{1}{p}[E(F^2(T_1)) - \{E(F(T_1))\}^2] \\
&= \frac{1}{p}\int F^2(T) dF_1(t) - \frac{1}{p}\left(\frac{1}{2} + O(\frac{1}{\sqrt{p}})\right)^2 \\
&= \frac{1}{p}\left(1 - \frac{1}{\sqrt{p}}\right)\int F^2(t) dF(t) + \frac{1}{p\sqrt{p}}\int F^2(t) dG(t) \\
&- \frac{1}{p}\left(\frac{1}{2} + O(\frac{1}{\sqrt{p}})\right)^2 \\
&= \frac{1}{p}\left(1 - \frac{1}{\sqrt{p}}\right)\frac{1}{3} + O\left(\frac{1}{p\sqrt{p}}\right) - \frac{1}{4p} \\
&= \frac{1}{12p} + O\left(\frac{1}{p\sqrt{p}}\right). \qquad (2.4)
\end{aligned}
$$

By using (2.4) and the result from A_1, we can obtain

$$
\mathrm{Var}(S_{N,p}) = E[\mathrm{Var}(S_{N,p})|T_1, \ldots, T_p] + \mathrm{Var}[E(S_{N,p}|T_1, \ldots, T_p)]
$$

$$= \frac{1}{12N} + \frac{1}{12p} + O\left(\frac{1}{p\sqrt{p}}\right) + O\left(\frac{1}{N\sqrt{p}}\right) \tag{2.5}$$

Hence, $E\left[\left(S_{N,P} - U_p - V_N + \frac{1}{2}\right)^2\right]$ is of smaller order than $\mathrm{Var}(S_{N,p})$.

Corollary 2.6.2 *If* $\frac{p}{N} \to c$, *where* $c > 0$ *is a constant as* $N \to \infty$ *and* $p \to \infty$, *then* $\frac{\sqrt{p}\left(\sum_{i=1}^{p}(1 - \hat{F}_N(T_i)) - \frac{1}{2}\right)}{\sqrt{1/12}}$ *converges to a normal distribution with mean* $\tilde{\mu}$ *and variance* $1 + c$ *as* $N \to \infty$ *and* $p \to \infty$, *where* $\tilde{\mu} = \frac{\int(1 - F(t))dG(t) - \frac{1}{2}}{\sqrt{\frac{1}{12}(1+c)}}$.

Proof Let $\sigma^2_{N,p}$ in (2.5) be the variance of $S_{N,p}$. Let σ^2_p and σ^2_N be the variance of U_p and V_N, respectively. Note that σ^2_p was found in (2.4), and σ^2_N is trivially $(12N)^{-1}$. The random variables $\frac{S_{N,p} - \frac{1}{2}}{\sigma_{N,p}}$ and $\frac{U_p - \frac{1}{2}}{\sigma_{N,p}} + \frac{V_N - \frac{1}{2}}{\sigma_{N,p}}$ have the same asymptotic distribution by Theorem 2.6.1. We can easily show that $\frac{\sigma_p}{\sigma_{N,p}}$ and $\frac{\sigma_N}{\sigma_{N,p}}$ converge to $\frac{1}{\sqrt{1+c}}$ and $\frac{\sqrt{c}}{\sqrt{1+c}}$, respectively, and $\frac{U_p - \frac{1}{2}}{\sigma_{N,p}}$ can be written as

$$\frac{U_p - E(1 - F(T))}{\sigma_p}\frac{\sigma_p}{\sigma_{N,p}} + \frac{E(1 - F(T)) - \frac{1}{2}}{\sigma_{N,p}}.$$

It can be shown that $\frac{E(1 - F(T)) - \frac{1}{2}}{\sigma_{N,p}}$ converges to $\frac{\int(1 - F(t))dG(t) - \frac{1}{2}}{\sqrt{\frac{1}{12}(1+c)}}$, by using the fact $E(1 - F(T)) = \left(1 - \frac{1}{\sqrt{p}}\right)\frac{1}{2} + \frac{1}{\sqrt{p}}\int(1 - F(t))dG(t)$. Hence, by Slutsky's theorem, $\frac{U_p - \frac{1}{2}}{\sigma_{N,p}}$ converges to a normal distribution with mean $\tilde{\mu}$ and variance $\frac{1}{1+c}$. Similarly, we can show that $\frac{V_N - \frac{1}{2}}{\sigma_{N,p}}$ converges to a normal distribution with mean 0 and variance $\frac{c}{1+c}$. Since $\frac{U_p - \frac{1}{2}}{\sigma_{N,p}}$ and $\frac{V_N - \frac{1}{2}}{\sigma_{N,p}}$ are independent, $\frac{S_{N,p} - \frac{1}{2}}{\sigma_{N,p}}$ converges to a normal distribution with mean $\tilde{\mu}$ and standard deviation 1. By using the fact that $\frac{\sigma_{N,p}}{\sqrt{1/12p}}$ converges to $\sqrt{1+c}$, we can prove the result through applying Slutsky's

theorem.

Corollary 2.6.2 implies that $\sqrt{p}$-alternatives can be detected using Edgington's method based on the empirical P-values, as long as $\int (1 - F(t))dG(t)$ is less than $1/2$ and we do the one-sided test. This result makes sense because we expect that the expectation of a P-value under the alternative is less than $1/2$ when the gof test is not biased.

3. SIMULATIONS

In the simulation study, three null distributions, normal, Laplace and Weibull, are considered. Since the exact null distribution of AD, CvM or Watson is unknown, 100,000 bootstrap replications were used to obtain empirical P-values. The empirical power and size presented in this section are obtained from 2,000 replications at the significance level α=0.05.

3.1 Testing whether data come from normal distributions

The normal distribution is one of the most widely used and important distributions in statistics. Its popularity comes from both the central limit theorem and the fact that many natural phenomena, such as height and lengths of items produced from machines, follow normal distributions. Also, many simple statistical methods like the t-test, linear discriminant analysis and analysis of variance assume normality. Hence, it is often essential to verify that data sets come from normal distributions, especially when we have data sets with few observations, because we cannot use the central limit theorem in this case. In our simulations, two alternative distributions, t-distribution with 10 degrees of freedom and chi-squared distribution with 10 degrees of freedom are considered since these are relatively close to normality and hence difficult to detect. Two alternative distributions which are further from normality, Cauchy, and Laplace distributions, are also selected. Since AD, CvM and Watson use a distance between F_n, the empirical cdf, and $\Phi\left(\frac{x-\mu}{\sigma}\right)$, the theoretical cdf under the null, we need to estimate location and scale parameters, μ and σ, and these are estimated by maximum likelihood estimators (MLE) in our simulations.

Tables 3.1, 3.3 and 3.5 show the empirical power and size of the one-sided moment based tests and smoothing based tests when every data set comes from the same

distribution. Of the two moment based tests, Fisher's method has higher power than Edgington's method, and smoothing based tests perform similarly to each other. For all considered alternatives, there exists no bias problem. This implies that the power would decrease when the two-sided test is applied. Tables 3.2, 3.4 and 3.6 show the power of two-sided tests when significance level α_2=0.01 is used. From these tables, we notice that the power of the two-sided moment based tests tends to be between the power of the one-sided moment based tests and that of smoothing based tests, showing little loss in power.

Since both the power itself and the relative decrease in power depend on the significance levels α_2, the effects of the significance levels α_2 are investigated. The relative decrease in power is defined as the ratio of the power of two-sided moment based tests subtracted from that of the one-sided moment based test to the power of the one-sided moment based test, and it has positive values when power decreases as the result of applying the two-sided tests. Figures 3.1, 3.2 and 3.3 show the power and the relative power decrease as a function of the significance level α_2 when we have 100 data sets with 5 observations. This case is selected because the effects of the two-sided tests might be more severe than in other cases. Since the power is always 100% when the alternative is a Cauchy distribution, this alternative is not considered. In each figure, the left and right plots show changes and the relative decrease in power as a function of α_2. From these plots, we notice that when there is no bias problem, the relative decrease in power increases as the significance level α_2 increases. Except for Laplace alternatives, the power tends to decrease more than 30% when evenly divided significance levels are used. When the alternative is the t-distribution or a Laplace distribution, the power decreases more than when Edgington's method is used. The amount of decrease in the power is similar for both methods under the chi-squared distribution.

Tables 3.7 to 3.12 show the local power of tests, i.e., 90% of data sets come from normal distributions and the remaining data sets are from alternative distributions. When the alternative is the t-distribution, the power of both moment based tests and smoothing based tests is just a little bit above the significance level, 0.05. This may not be a surprising result because t-distributions with large degrees of freedom are close to normal distributions. Effects of the two-sided tests are investigated in Figures 3.4 to 3.7. When the alternative is a Cauchy distribution, Edgington's method shows more decrease in the power. In Figure 3.4, we notice that the power of the two-sided tests is below the size of the test when the significance level α_2 is greater than 0.018. This indicates that selecting the two significance levels is crucial, and both the figure and the results in Section 2.3. suggest that it might be best to use the significance level α_2 less than 0.015.

Under local alternatives, we notice that Fisher's method tends to have higher power than Edgington's method. The difference in the power between these two methods tends to be large when the alternative is a Cauchy or a Laplace distribution. The reason can be explained by the density of P-values. Figure 3.8 shows the density of the P-value when the sample size is 5 and CvM is applied to every small data set. When the alternative is a Laplace distribution, there is stronger evidence against the null. The density of the P-value under Cauchy distributions, which is not shown here, exhibits much stronger evidence against the null. In Section 2.3. we found that Fisher's method is asymptotically more powerful than Edgington's method when there exists stronger evidence against the null. Both higher power of Fisher's method and the density of the P-value support findings in this section.

We might obtain more insights about the performance of test procedures if the power is investigated when several different proportions of data sets are from the null. Only two alternatives, the t-distribution, and the chi-squared distribution are

selected for illustration purposes. Since results for CvM, Watson and AD are similar, only the results for AD are shown here.

Figures 3.9 and 3.10 show the empirical power when the alternative is a mixture of normal and t-distributions. In the plots, the proportion denotes the proportion of data sets which are from the null distribution, and Figures 3.9 and 3.10 represent the empirical power when the sample sizes are 5 and 10, respectively. Regardless of the sample size, moment based tests dominate smoothing based tests. Especially, Fisher's method has better power than Edgington's method. Also, under both sample sizes, the power of smoothing based tests is just around the size when at least 70% of data sets are from the null. On the contrary, the power of moment based tests when more than 70% of data sets are from the null is above the size.

Figures 3.11 and 3.12 show the empirical power when the alternative is a mixture of normal and chi-squared distributions. From Figure 3.11, we notice that Fisher's method is the best regardless of the considered number of data sets when the sample size is 5. However, when the sample size is 10, it is hard to tell which method is the best. Especially, when less than 50% of data sets are from the null and there are 500 or 1,000 data sets, the power of each method is approximately 1, indicating little effect due to the method of combining P-values. We still notice, however, that Fisher's method outperforms the others when at least 50% of data sets are from the null and there are 100 or 300 data sets.

When testing normality, moment based tests are generally better than smoothing based tests, and Fisher's method has higher power than Edgington's method. We can see little difference between the power of the three tests, AD, CvM, and Watson. Hence, we might conclude from the simulation results that when we test normality, moment based tests using P-values from AD, CvM or Watson are more desirable than smoothing based tests.

Table 3.1: This table shows the size(%) and power(%) of the test. The null hypothesis is that data come from normal distributions and AD is applied to every small data set. For moment based tests, the one-sided test is used.

		Edgington's method					Fisher's method				
n	p	Normal	t(10)	$\chi^2(10)$	Laplace	Cauchy	Normal	t(10)	$\chi^2(10)$	Laplace	Cauchy
5	100	5.5	9.6	27.5	43.1	100.0	6.4	11.6	31.4	59.7	100.0
	300	5.2	14.0	55.5	80.8	100.0	5.5	17.8	62.2	94.5	100.0
	500	4.5	15.3	74.4	93.5	100.0	5.1	22.1	80.6	99.2	100.0
	1000	5.8	23.9	94.7	99.7	100.0	6.2	33.3	97.1	100.0	100.0
10	100	5.4	20.3	86.1	97.5	100.0	5.1	30.9	92.6	99.9	100.0
	300	3.8	40.6	100.0	100.0	100.0	4.1	60.8	100.0	100.0	100.0
	500	5.6	53.4	100.0	100.0	100.0	4.4	78.7	100.0	100.0	100.0
	1000	5.1	82.5	100.0	100.0	100.0	4.3	95.8	100.0	100.0	100.0
		Smooth Test					Order Selection Test				
n	p	Normal	t(10)	$\chi^2(10)$	Laplace	Cauchy	Normal	t(10)	$\chi^2(10)$	Laplace	Cauchy
5	100	5.0	6.4	14.7	29.6	100.0	5.3	5.5	16.8	29.6	100.0
	300	4.8	9.5	38.1	75.3	100.0	4.3	8.6	39.0	70.5	100.0
	500	4.8	9.0	59.1	92.0	100.0	4.8	8.5	57.6	88.5	100.0
	1000	5.5	15.8	89.3	99.9	100.0	5.5	14.6	87.2	99.5	100.0
10	100	4.4	12.6	72.7	96.1	100.0	4.4	12.1	74.7	94.3	100.0
	300	4.0	31.4	99.8	100.0	100.0	3.9	28.0	99.7	100.0	100.0
	500	4.3	45.1	100.0	100.0	100.0	4.0	39.6	100.0	100.0	100.0
	1000	4.8	76.3	100.0	100.0	100.0	4.5	72.3	100.0	100.0	100.0

Table 3.2: This table shows the size(%) and power(%) of the two-sided moment based test. The null hypothesis is that data come from normal distributions and AD is applied to every small data set. The significance level α_2=0.01 is used.

		Edgington's method					Fisher's method				
n	p	Normal	t(10)	$\chi^2(10)$	Laplace	Cauchy	Normal	t(10)	$\chi^2(10)$	Laplace	Cauchy
5	100	5.1	8.2	23.9	39.4	100.0	6.0	10.7	27.1	56.0	100.0
	300	4.8	11.9	51.0	77.6	100.0	5.2	15.3	58.2	93.1	100.0
	500	5.0	13.0	70.5	92.3	100.0	5.0	18.9	77.8	99.1	100.0
	1000	5.6	21.6	93.2	99.7	100.0	6.3	30.1	96.3	100.0	100.0
10	100	4.9	17.9	83.5	96.7	100.0	5.1	26.9	91.5	99.8	100.0
	300	3.9	36.6	99.9	100.0	100.0	4.1	57.4	100.0	100.0	100.0
	500	5.1	48.8	100.0	100.0	100.0	4.1	75.8	100.0	100.0	100.0
	1000	5.0	79.3	100.0	100.0	100.0	4.6	95.2	100.0	100.0	100.0

Table 3.3: This table shows the size(%) and power(%) of the test. The null hypothesis is that data come from normal distributions and CvM is applied to every small data set. For moment based tests, the one-sided test is used.

		Edgington's method					Fisher's method				
n	p	Normal	t(10)	$\chi^2(10)$	Laplace	Cauchy	Normal	t(10)	$\chi^2(10)$	Laplace	Cauchy
5	100	5.4	10.2	25.4	46.6	100.0	6.3	12.0	29.0	61.9	100.0
	300	5.4	14.6	51.9	83.8	100.0	5.4	18.2	59.3	95.6	100.0
	500	4.9	16.9	70.4	95.5	100.0	5.4	23.2	77.6	99.6	100.0
	1000	5.8	25.8	92.8	100.0	100.0	6.4	35.8	96.0	100.0	100.0
10	100	5.3	18.9	80.8	97.4	100.0	5.2	26.5	89.4	99.9	100.0
	300	4.2	37.5	99.9	100.0	100.0	4.1	55.5	99.9	100.0	100.0
	500	5.0	50.2	100.0	100.0	100.0	4.2	73.8	100.0	100.0	100.0
	1000	4.9	78.3	100.0	100.0	100.0	4.6	93.9	100.0	100.0	100.0
		Smooth Test					Order Selection Test				
n	p	Normal	t(10)	$\chi^2(10)$	Laplace	Cauchy	Normal	t(10)	$\chi^2(10)$	Laplace	Cauchy
5	100	4.5	7.2	14.1	32.3	100.0	5.0	6.6	15.5	32.3	100.0
	300	4.6	9.3	35.8	79.1	100.0	5.0	8.3	35.9	74.9	100.0
	500	4.7	10.3	55.1	93.7	100.0	5.0	9.2	53.8	91.0	100.0
	1000	5.5	17.1	85.5	99.9	100.0	5.5	15.8	83.1	99.9	100.0
10	100	4.2	11.0	64.7	96.2	100.0	3.9	10.5	67.8	94.8	100.0
	300	4.3	27.4	99.3	100.0	100.0	4.0	25.4	99.1	100.0	100.0
	500	4.2	40.7	100.0	100.0	100.0	4.0	36.2	100.0	100.0	100.0
	1000	4.7	70.5	100.0	100.0	100.0	4.4	66.9	100.0	100.0	100.0

Table 3.4: This table shows the size(%) and power(%) of the two-sided moment based test. The null hypothesis is that data come from normal distributions and CvM is applied to every small data set. The significance level α_2=0.01 is used.

		Edgington's method					Fisher's method				
n	p	Normal	t(10)	$\chi^2(10)$	Laplace	Cauchy	Normal	t(10)	$\chi^2(10)$	Laplace	Cauchy
5	100	4.7	7.7	21.3	40.6	100.0	5.3	10.3	24.9	57.5	100.0
	300	4.5	11.8	46.0	80.5	100.0	5.0	14.9	53.7	94.2	100.0
	500	4.8	13.3	65.3	94.3	100.0	4.6	18.2	72.3	99.3	100.0
	1000	5.2	22.1	90.3	99.8	100.0	5.9	29.8	94.6	100.0	100.0
10	100	4.8	15.4	76.4	96.5	100.0	4.7	22.6	86.8	99.9	100.0
	300	4.0	32.6	99.6	100.0	100.0	4.1	49.8	99.8	100.0	100.0
	500	4.6	44.8	100.0	100.0	100.0	3.7	69.0	100.0	100.0	100.0
	1000	4.4	73.2	100.0	100.0	100.0	4.2	92.0	100.0	100.0	100.0

Table 3.5: This table shows the size(%) and power(%) of the test. The null hypothesis is that data come from normal distributions and Watson is applied to every small data set. For moment based tests, the one-sided test is used.

		Edgington's method					Fisher's method				
n	p	Normal	t(10)	$\chi^2(10)$	Laplace	Cauchy	Normal	t(10)	$\chi^2(10)$	Laplace	Cauchy
5	100	5.3	9.4	24.8	44.0	100.0	6.4	11.9	29.3	60.2	100.0
	300	5.1	13.8	50.0	81.5	100.0	5.6	18.1	59.1	95.0	100.0
	500	4.8	14.8	68.2	94.2	100.0	5.5	22.9	77.5	99.4	100.0
	1000	5.5	23.4	91.7	99.8	100.0	7.0	34.1	96.1	100.0	100.0
10	100	5.1	16.3	75.9	96.5	100.0	5.5	24.1	86.6	99.8	100.0
	300	3.8	32.3	99.2	100.0	100.0	4.3	50.1	99.8	100.0	100.0
	500	4.2	42.2	100.0	100.0	100.0	4.5	68.0	100.0	100.0	100.0
	1000	3.8	67.7	100.0	100.0	100.0	5.2	90.8	100.0	100.0	100.0
		Smooth Test					Order Selection Test				
n	p	Normal	t(10)	$\chi^2(10)$	Laplace	Cauchy	Normal	t(10)	$\chi^2(10)$	Laplace	Cauchy
5	100	4.8	7.3	13.9	29.8	100.0	5.2	6.2	14.9	29.5	100.0
	300	5.1	8.8	34.7	75.8	100.0	4.8	8.2	34.5	72.0	100.0
	500	5.0	9.6	53.1	91.8	100.0	5.1	8.6	51.9	89.1	100.0
	1000	5.4	14.8	83.8	99.8	100.0	5.5	13.9	81.8	99.6	100.0
10	100	3.8	9.6	58.6	95.2	100.0	4.2	8.6	61.3	92.9	100.0
	300	4.2	22.0	98.4	100.0	100.0	4.3	19.9	98.0	100.0	100.0
	500	4.3	32.6	99.9	100.0	100.0	4.2	28.9	99.9	100.0	100.0
	1000	4.6	58.6	100.0	100.0	100.0	4.6	53.8	100.0	100.0	100.0

Table 3.6: This table shows the size(%) and power(%) of the two-sided moment based test. The null hypothesis is that data come from normal distributions and Watson is applied to every small data set. The significance level α_2=0.01 is used.

		Edgington's method					Fisher's method				
n	p	Normal	t(10)	$\chi^2(10)$	Laplace	Cauchy	Normal	t(10)	$\chi^2(10)$	Laplace	Cauchy
5	100	4.6	7.3	20.3	37.9	100.0	5.6	10.5	24.4	55.2	100.0
	300	4.5	10.9	44.4	77.6	100.0	4.9	14.7	53.3	93.3	100.0
	500	4.6	11.6	62.5	92.2	100.0	4.7	18.2	72.0	99.1	100.0
	1000	5.0	19.4	88.2	99.5	100.0	6.2	29.1	94.5	100.0	100.0
10	100	4.6	13.0	71.2	95.4	100.0	5.1	20.7	83.5	99.8	100.0
	300	3.8	27.3	98.9	100.0	100.0	4.3	44.1	99.8	100.0	100.0
	500	3.8	36.8	100.0	100.0	100.0	4.0	62.7	100.0	100.0	100.0
	1000	3.8	61.9	100.0	100.0	100.0	4.8	88.0	100.0	100.0	100.0

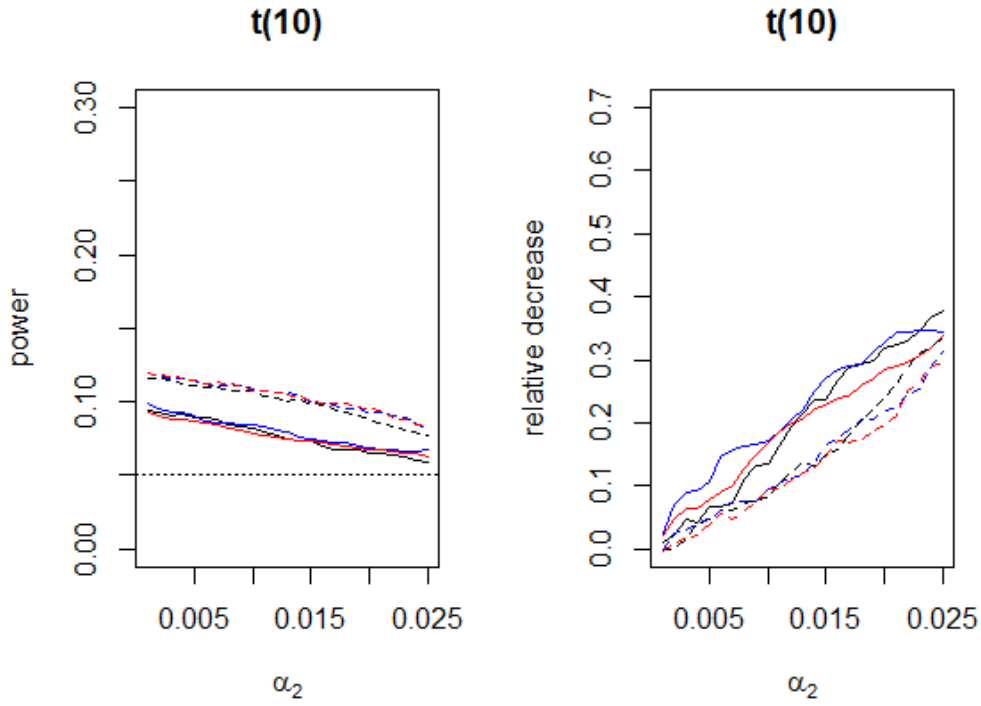

Figure 3.1: The left and right plots show the power of the two-sided moment based tests and the relative power decrease over various significance levels α_2, when there are 100 data sets with 5 observations. In both plots, black, blue and red lines represent AD, CvM, and Watson, respectively. Also, the solid and dashed lines represent Edgington's method and Fisher's method, respectively. The dotted line in the left plot denotes the significance level α=0.05.

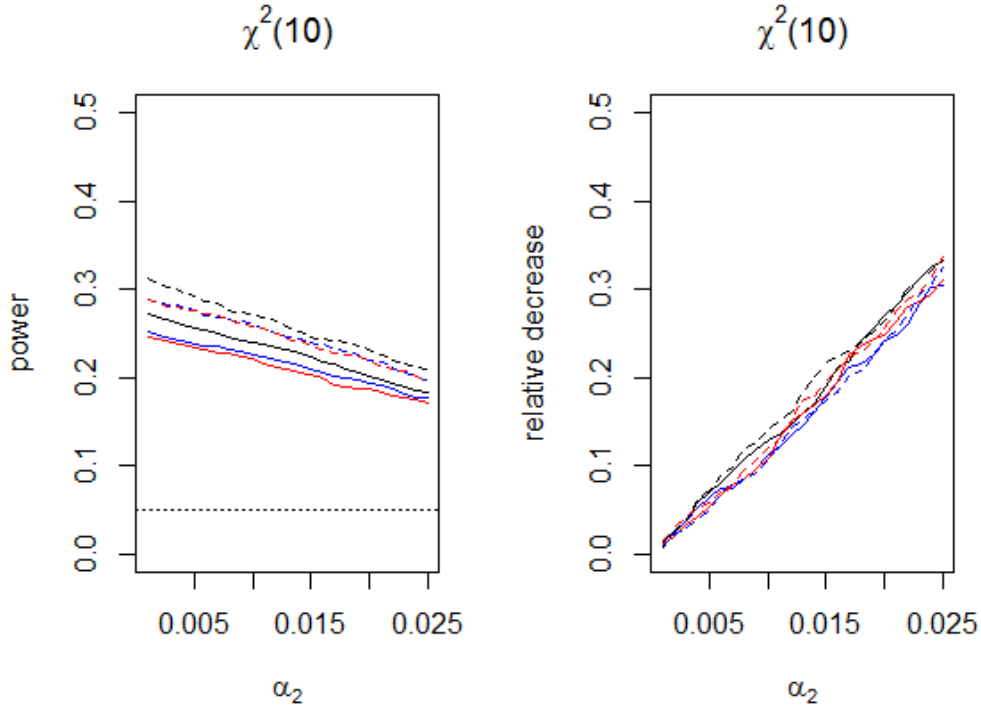

Figure 3.2: The left and right plots show the power of the two-sided moment based tests and the relative power decrease over various significance levels α_2, when there are 100 data sets with 5 observations. In both plots, black, blue and red lines represent AD, CvM, and Watson, respectively. Also, solid and dashed lines represent Edgington's method and Fisher's method, respectively. The dotted line in the left plot denotes the significance level α=0.05.

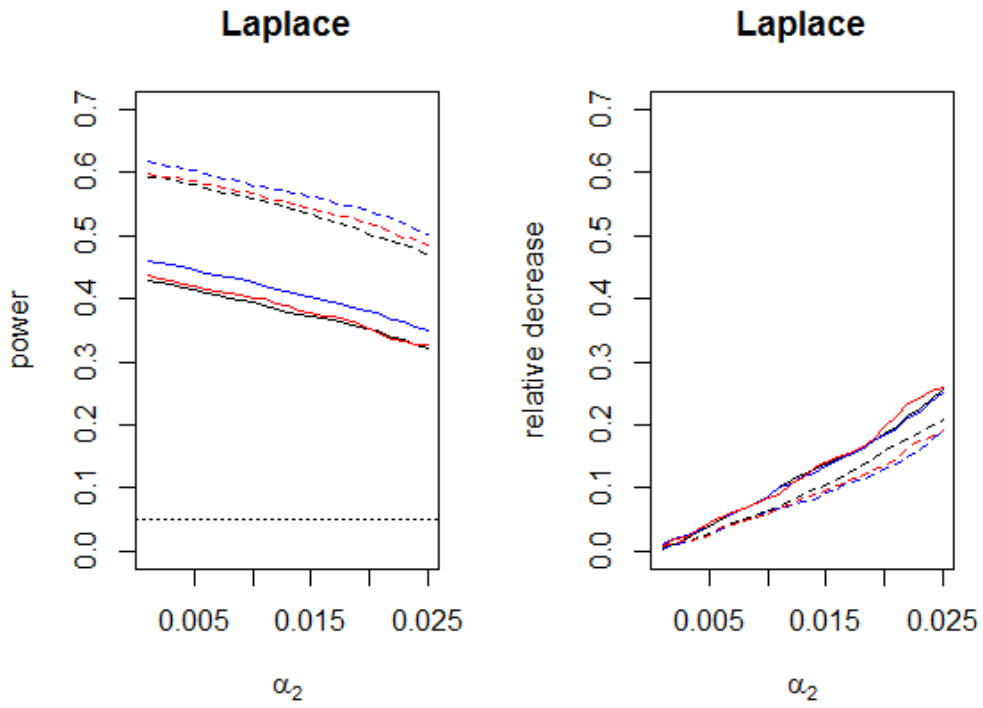

Figure 3.3: The left and right plots show the power of the two-sided tests and the relative power decrease over various significance levels α_2, when there are 100 data sets with 5 observations. In both plots, black, blue and red lines represent AD, CvM, and Watson, respectively. Also, the solid and dashed lines represent Edgington's method and Fisher's method, respectively. The dotted line in the left plot denotes the significance level α=0.05.

Table 3.7: This table shows the local power(%) of the test when 90% of data sets are from the null distribution. The null hypothesis is that data come from normal distributions and AD is applied to every small data set. For moment based tests, the one-sided test is used.

		Edgington's method				Fisher's method			
n	p	t(10)	$\chi^2(10)$	Laplace	Cauchy	t(10)	$\chi^2(10)$	Laplace	Cauchy
5	100	6.2	6.3	6.6	15.8	6.1	7.0	7.6	42.6
	300	6.0	7.8	8.3	33.2	6.6	7.5	10.5	79.8
	500	6.2	8.2	9.6	42.9	5.6	8.9	12.8	92.6
	1000	7.3	9.9	13.5	67.2	7.4	11.6	17.6	99.4
10	100	5.9	8.5	10.8	34.1	6.1	9.5	16.7	96.0
	300	6.2	13.2	15.2	69.7	7.8	16.1	27.2	100.0
	500	8.2	15.6	21.2	86.1	9.1	21.9	36.2	100.0
	1000	9.9	25.1	31.5	98.2	11.1	33.9	54.6	100.0
		Smooth Test				Order Selection Test			
n	p	t(10)	$\chi^2(10)$	Laplace	Cauchy	t(10)	$\chi^2(10)$	Laplace	Cauchy
5	100	5.5	4.8	5.1	15.3	4.5	5.1	4.9	8.8
	300	5.1	5.5	5.1	29.7	5.3	5.3	5.5	20.5
	500	4.6	5.3	6.0	44.0	5.0	5.2	5.8	30.5
	1000	4.8	7.1	7.8	73.4	5.0	6.8	7.5	57.4
10	100	4.9	5.5	7.1	59.4	4.8	5.5	6.6	22.1
	300	4.9	8.8	10.9	95.0	5.2	8.6	8.8	63.0
	500	6.2	9.7	14.6	99.9	6.4	9.8	13.1	89.2
	1000	6.8	16.8	22.2	100.0	7.0	15.6	19.1	100.0

Table 3.8: This table shows the local power(%) of the two-sided moment based test when 90% of data sets are from the null distribution. The null hypothesis is that data come from normal distributions and AD is applied to every small data sets. The significance level α_2=0.01 is used.

		Edgington's method				Fisher's method			
n	p	t(10)	$\chi^2(10)$	Laplace	Cauchy	t(10)	$\chi^2(10)$	Laplace	Cauchy
5	100	5.2	5.5	5.8	12.9	5.1	6.3	7.1	38.6
	300	5.8	6.6	7.1	29.0	6.1	6.8	8.7	77.3
	500	6.0	7.2	8.1	38.9	5.0	7.8	10.6	90.8
	1000	6.6	9.3	11.8	63.2	6.9	9.8	15.4	99.4
10	100	5.3	7.6	8.8	29.8	5.3	8.2	14.0	95.4
	300	5.7	11.3	13.2	65.9	7.1	14.2	23.9	100.0
	500	7.8	13.2	19.1	83.2	7.8	18.8	32.4	100.0
	1000	8.5	21.7	27.3	97.9	9.2	29.6	50.6	100.0

Table 3.9: This table shows the local power(%) of the test when 90% of data sets are from the null distribution. The null hypothesis is that data come from normal distributions and CvM is applied to every small data set. For moment based tests, the one-sided test is used.

		Edgington's method				Fisher's method			
n	p	t(10)	$\chi^2(10)$	Laplace	Cauchy	t(10)	$\chi^2(10)$	Laplace	Cauchy
5	100	6.4	6.2	6.8	16.3	5.9	6.6	8.1	42.8
	300	6.2	8.0	8.3	34.2	6.3	7.6	10.7	80.5
	500	6.6	8.6	10.4	45.4	5.8	8.5	13.0	92.8
	1000	7.7	10.6	15.6	70.2	7.4	11.2	18.2	99.5
10	100	5.7	8.5	11.3	33.7	6.2	9.0	16.2	96.2
	300	5.9	12.6	15.2	68.5	6.8	14.3	25.7	100.0
	500	8.3	14.2	22.1	85.5	8.2	18.6	35.7	100.0
	1000	10.3	22.6	31.6	98.6	9.6	29.1	53.3	100.0
		Smooth Test				Order Selection Test			
n	p	t(10)	$\chi^2(10)$	Laplace	Cauchy	t(10)	$\chi^2(10)$	Laplace	Cauchy
5	100	5.6	4.8	5.4	15.2	4.5	5.3	5.3	9.4
	300	5.2	5.3	5.2	30.9	5.8	5.3	5.2	20.9
	500	4.5	5.4	6.6	44.8	5.0	5.0	6.3	31.6
	1000	4.8	7.2	8.6	74.4	5.1	6.8	8.2	60.0
10	100	4.8	5.1	6.6	58.2	5.1	5.4	6.6	21.4
	300	5.1	7.5	10.6	95.3	5.2	7.4	9.2	62.6
	500	6.8	8.5	14.7	99.9	6.8	8.6	12.6	88.8
	1000	7.0	14.5	20.9	100.0	6.8	13.9	19.1	100.0

Table 3.10: This table shows the local power(%) of the two-sided moment based test when 90% of data sets are from the null distribution. The null hypothesis is that data come from normal distributions and CvM is applied to every small data set. The significance level α_2=0.01 is used.

		Edgington's method				Fisher's method			
n	p	t(10)	$\chi^2(10)$	Laplace	Cauchy	t(10)	$\chi^2(10)$	Laplace	Cauchy
5	100	5.7	5.4	6.3	13.3	5.1	5.9	7.4	38.9
	300	6.1	6.9	7.5	30.1	6.2	6.8	8.9	77.2
	500	6.0	7.0	8.8	40.6	5.0	7.3	11.1	90.8
	1000	6.7	9.3	13.2	65.8	6.6	9.1	16.2	99.4
10	100	5.4	7.5	9.4	29.5	5.5	8.0	14.1	95.8
	300	5.5	11.2	13.8	64.6	6.6	12.6	22.9	100.0
	500	7.4	11.8	18.8	83.7	7.4	16.2	31.4	100.0
	1000	8.7	20.0	28.3	97.9	8.6	25.6	49.4	100.0

Table 3.11: This table shows the local power(%) of the test when 90% of data sets are from the null distribution. The null hypothesis is that data come from normal distributions and Watson is applied to every small data set. For moment based tests, the one-sided test is used.

		Edgington's method				Fisher's method			
n	p	t(10)	$\chi^2(10)$	Laplace	Cauchy	t(10)	$\chi^2(10)$	Laplace	Cauchy
5	100	6.2	6.2	6.7	15.2	5.8	6.7	8.2	42.6
	300	5.2	7.0	9.1	34.2	5.9	7.9	11.1	79.3
	500	5.9	9.2	11.1	43.4	7.5	10.2	13.2	92.5
	1000	7.4	10.2	12.5	67.1	8.4	12.7	18.3	99.5
10	100	5.5	7.5	9.3	32.2	6.2	10.5	14.9	96.0
	300	6.4	10.6	14.7	62.8	8.1	14.7	25.8	100.0
	500	5.1	12.4	18.1	84.5	7.5	18.6	35.5	100.0
	1000	5.9	17.2	26.9	98.4	9.4	27.2	54.9	100.0
		Smooth Test				Order Selection Test			
n	p	t(10)	$\chi^2(10)$	Laplace	Cauchy	t(10)	$\chi^2(10)$	Laplace	Cauchy
5	100	5.6	4.8	5.4	15.2	4.5	5.3	5.3	9.4
	300	5.2	5.3	5.2	30.9	5.8	5.3	5.2	20.9
	500	4.5	5.4	6.6	44.8	5.0	5.0	6.3	31.6
	1000	4.8	7.2	8.6	74.4	5.1	6.8	8.2	60.0
10	100	4.8	5.1	6.6	58.2	5.1	5.4	6.6	21.4
	300	5.1	7.5	10.6	95.3	5.2	7.4	9.2	62.6
	500	6.8	8.5	14.7	99.9	6.8	8.6	12.6	88.8
	1000	7.0	14.5	20.9	100.0	6.8	13.9	19.1	100.0

Table 3.12: This table shows the local power(%) of the two-sided moment based test when 90% of data sets are from the null distribution. The null hypothesis is that data come from normal distributions and Watson is applied to every small data set. The significance level α_2=0.01 is used.

		Edgington's method				Fisher's method			
n	p	t(10)	$\chi^2(10)$	Laplace	Cauchy	t(10)	$\chi^2(10)$	Laplace	Cauchy
5	100	5.3	5.4	6.0	12.8	5.3	6.3	7.4	39.0
	300	5.3	6.2	7.7	29.6	6.2	7.2	9.0	76.9
	500	5.4	7.9	9.1	39.6	6.5	8.8	11.3	91.0
	1000	7.2	8.8	11.0	63.6	7.6	10.7	15.6	99.3
10	100	5.1	6.4	8.2	28.1	5.5	8.9	13.6	95.5
	300	6.2	8.7	12.9	59.0	7.0	13.0	22.4	100.0
	500	5.3	10.7	15.8	82.1	6.6	16.2	31.9	100.0
	1000	5.5	14.4	22.9	98.1	8.0	24.3	50.8	100.0

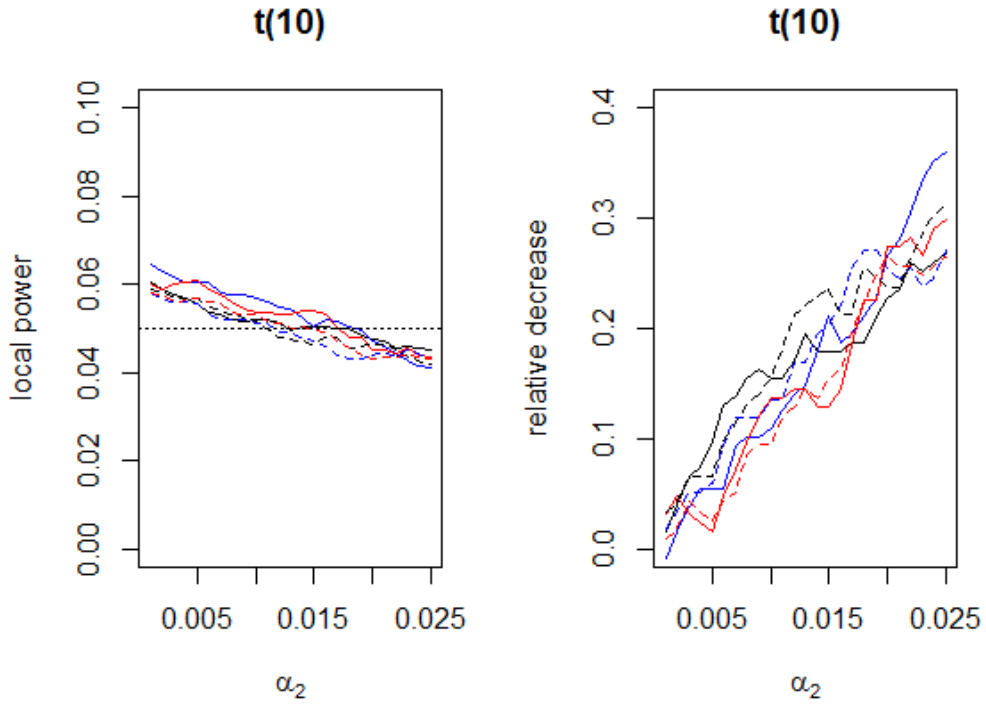

Figure 3.4: The left and right plots show the local power of two-sided moment based tests and the relative power decrease over various significance levels α_2, when there are 100 data sets with 5 observations. In both plots, black, blue and red lines represent AD, CvM, and Watson, respectively. Also, the solid and dashed lines represent Edgington's method and Fisher's method, respectively. The dotted line in the left plot denotes the significance level α=0.05.

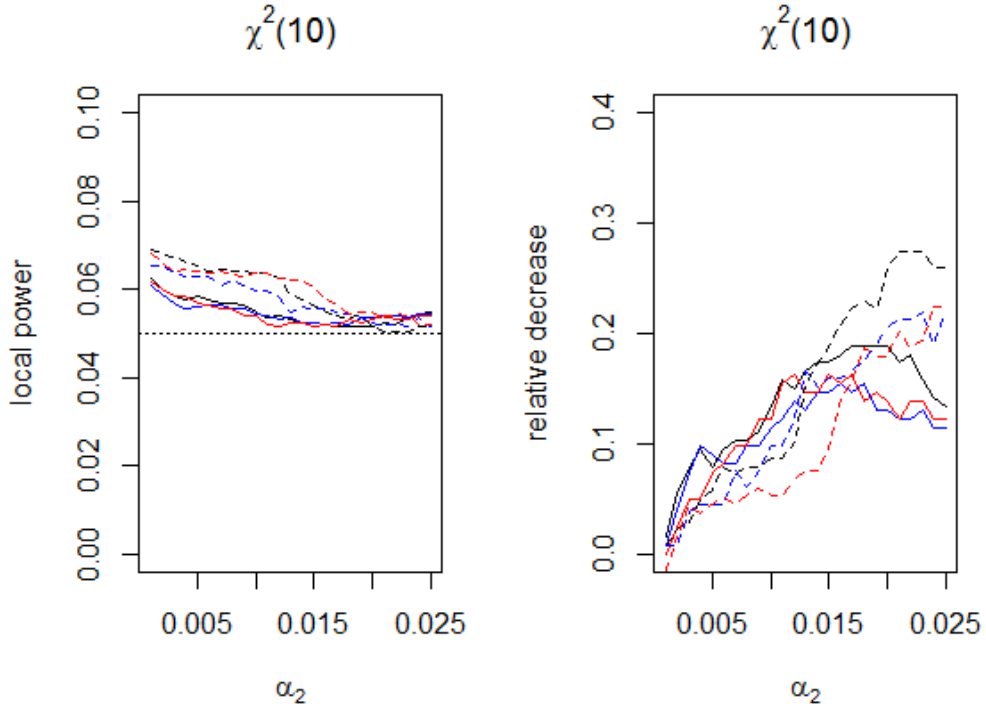

Figure 3.5: The left and right plots show the local power of the two-sided moment based tests and the relative power decrease over various significance levels α_2, when there are 100 data sets with 5 observations. In both plots, black, blue and red lines represent AD, CvM, and Watson, respectively. Also, solid and dashed lines represent Edgington's method and Fisher's method, respectively. The dotted line in the left plot denotes the significance level α=0.05.

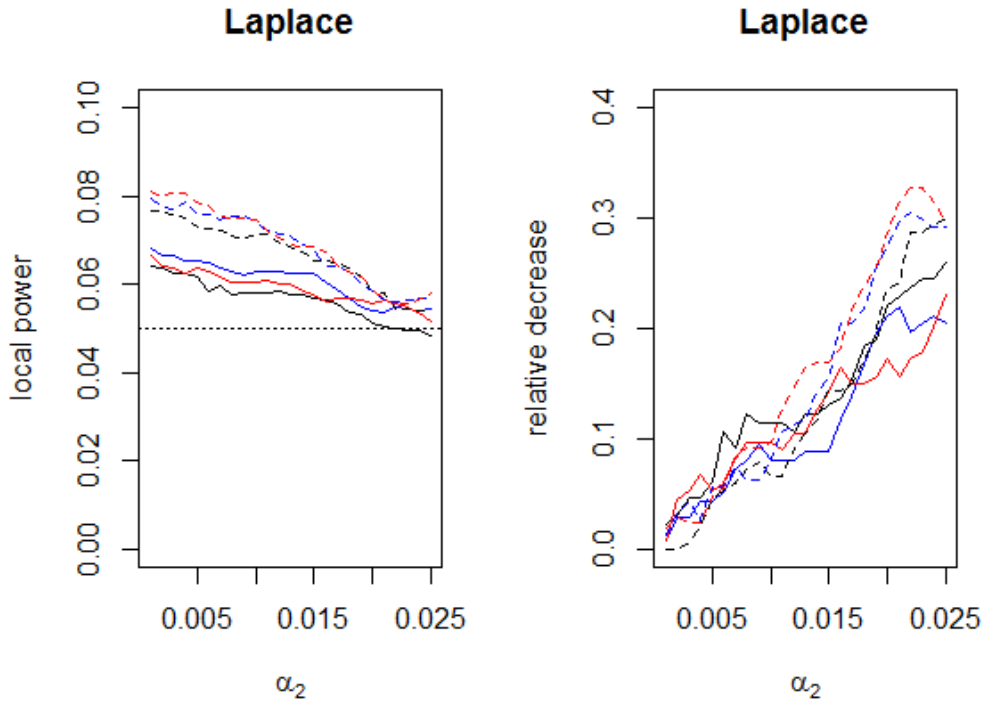

Figure 3.6: The left and right plots show the local power of the two-sided moment based tests and the relative power decrease over various significance levels α_2, when there are 100 data sets with 5 observations. In both plots, black, blue and red lines represent AD, CvM, and Watson, respectively. Also, the solid and dashed lines represent Edgington's method and Fisher's method, respectively. The dotted line in the left plot denotes the significance level α=0.05.

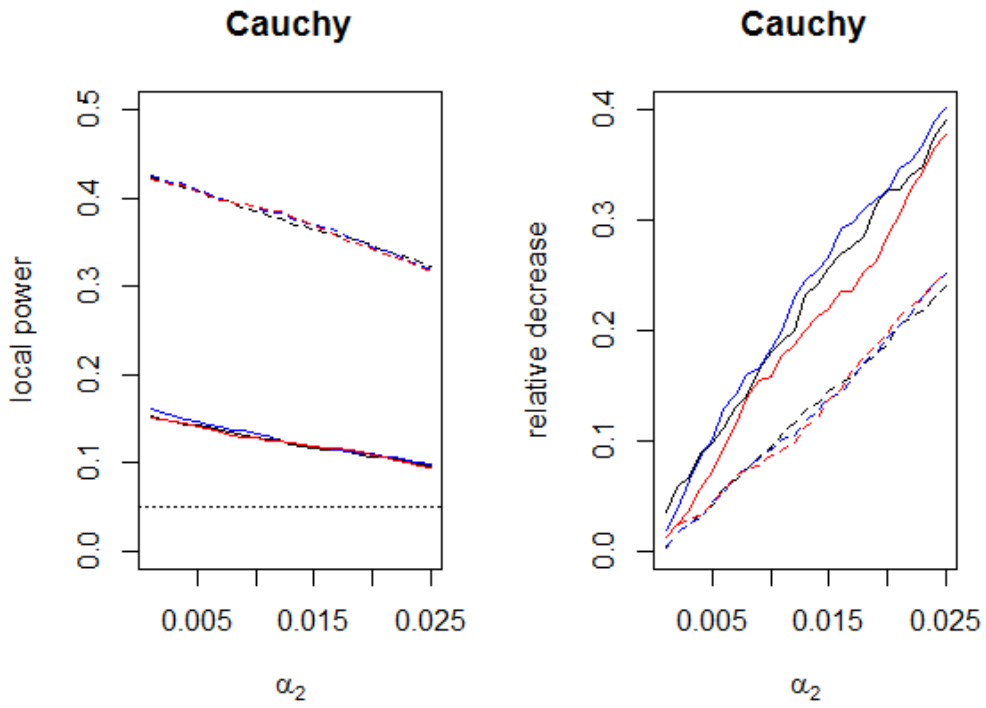

Figure 3.7: The left and right plots show the local power of the two-sided moment based tests and the relative power decrease over various significance levels α_2, when there are 100 data sets with 5 observations. In both plots, black, blue and red lines represent AD, CvM, and Watson, respectively. Also, the solid and dashed lines represent Edgington's method and Fisher's method, respectively. The dotted line in the left plot denotes the significance level α=0.05.

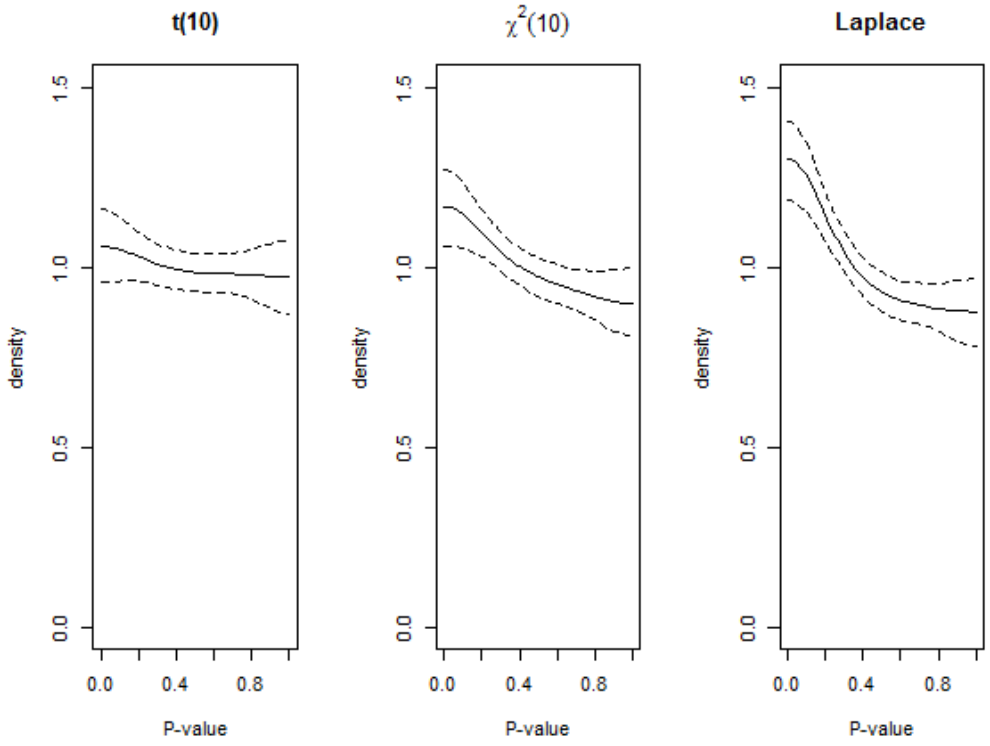

Figure 3.8: This figure shows the density of the P-value when CvM is applied to every small data set with sample sizes 5. The solid line is the median of 100 kernel density estimates and the dashed lines are 0.025 percentiles and 0.975 percentiles of kernel density estimates.

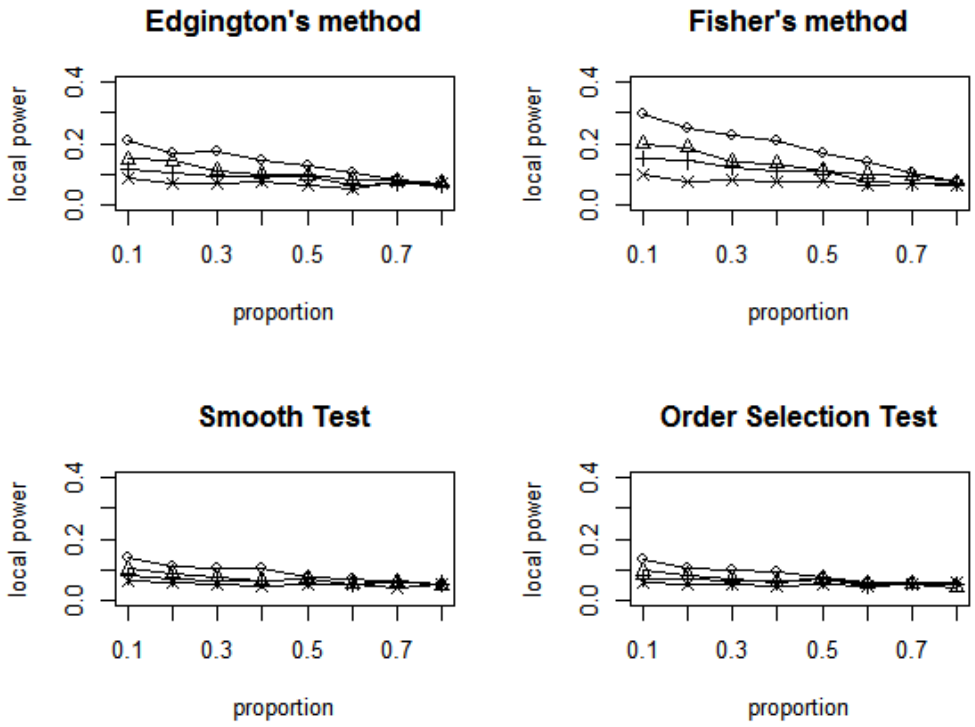

Figure 3.9: This figure shows the empirical power at the significance level 0.05 when testing whether data come from normal distributions, and the alternative is a mixture of normal and the t-distribution. The number of data sets considered are 100, 300, 500 and 1000, and cross, plus, triangle and circle represent the number of data sets, respectively. AD is applied to every small data with sample sizes 5.

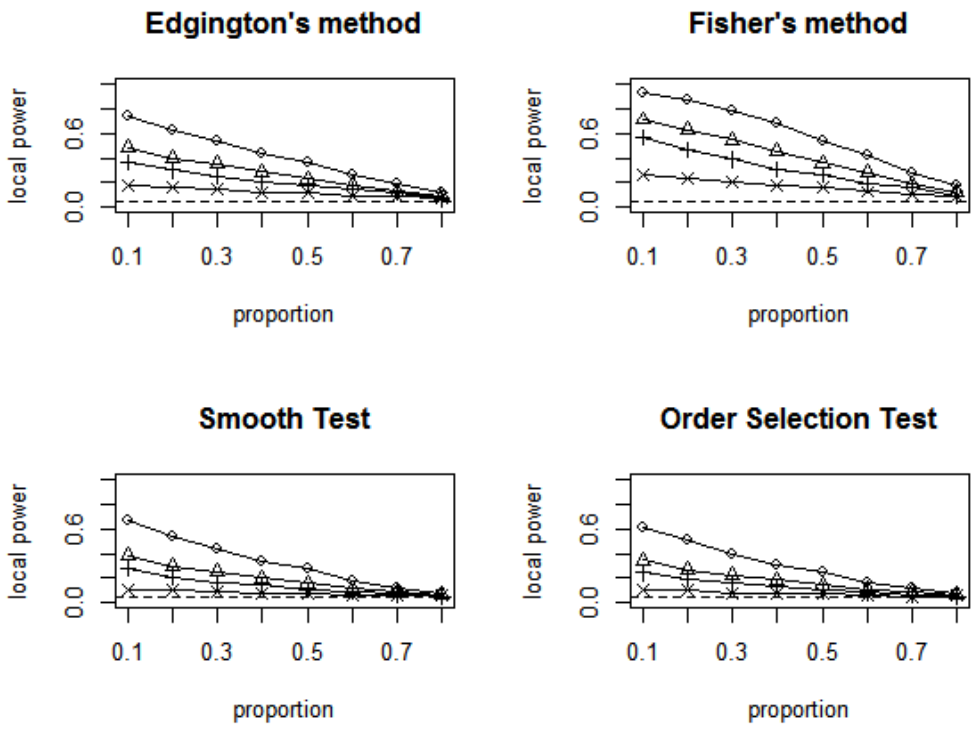

Figure 3.10: This figure shows the empirical power at the significance level 0.05 when testing whether data come from normal distributions, and the alternative is a mixture of normal and the t-distribution. The number of data sets considered are 100, 300, 500 and 1000, and cross, plus, triangle and circle represent the number of data sets, respectively. AD is applied to every small data set with sample sizes 10.

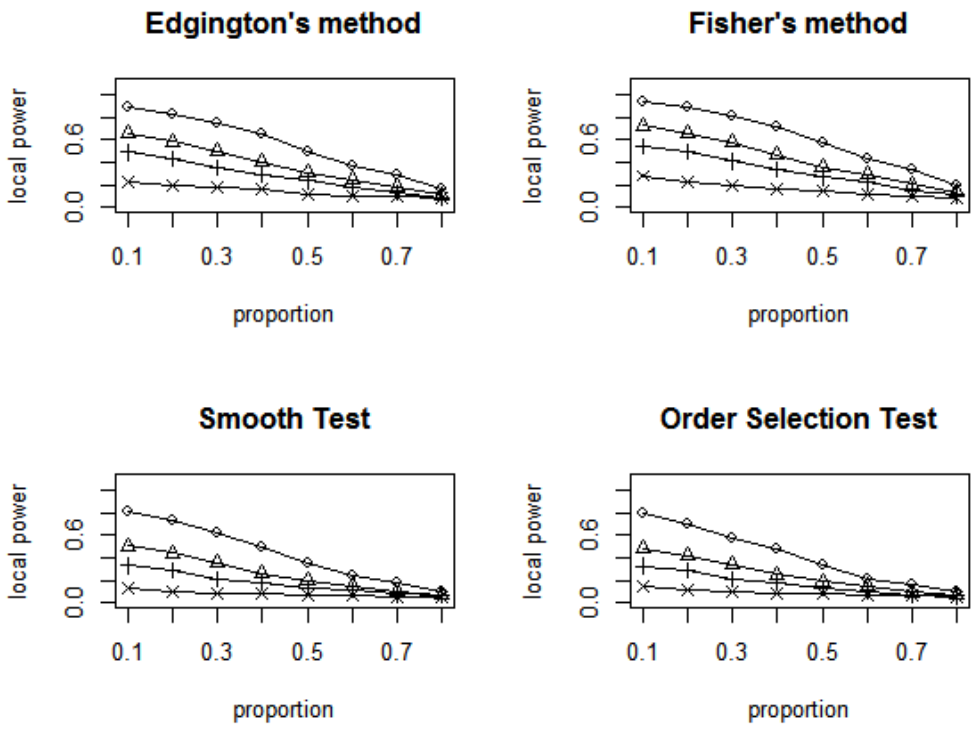

Figure 3.11: This figure shows the empirical power at the significance level 0.05 when testing whether data come from normal distributions, and the alternative is a mixture of normal and the chi-squared distribution. The number of data sets considered are 100,300, 500 and 1000, and cross, plus, triangle and circle represent the number of data sets, respectively. AD is applied to every small data set with sample sizes 5.

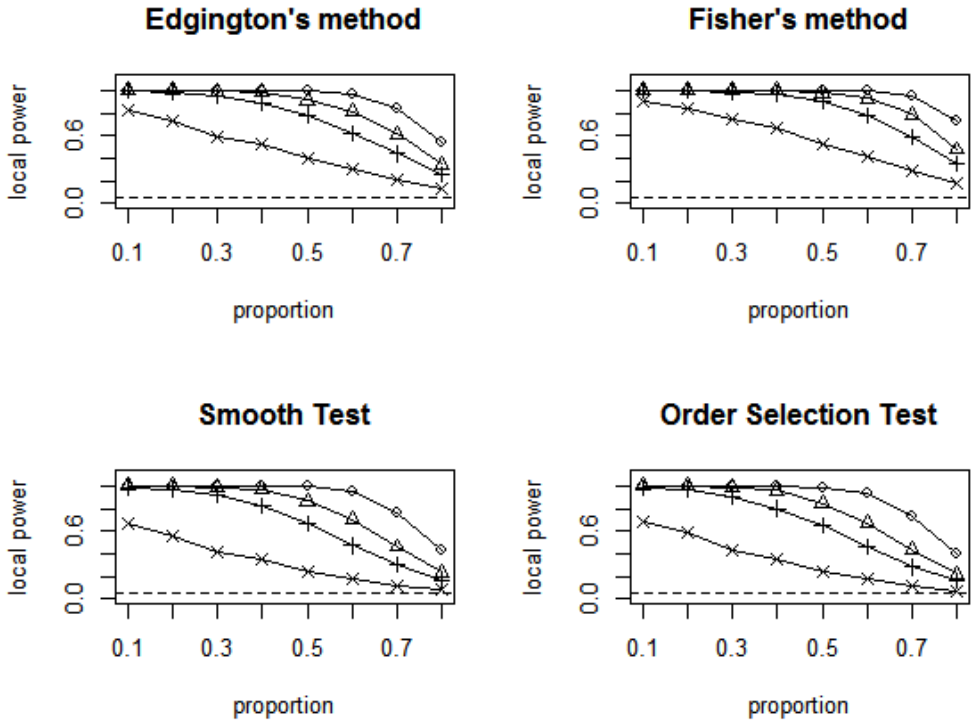

Figure 3.12: This figure shows the empirical power at the significance level 0.05 when testing whether data come from normal distributions, and the alternative is a mixture of normal and the chi-squared distribution. The number of considered data sets are 100, 300, 500 and 1000, and cross, plus, triangle and circle represent the number of data sets, respectively. AD is applied to every small data set with sample sizes 10.

3.2 Testing whether data come from Laplace distributions

The Laplace distribution, also known as the double exponential distribution, has been used in a variety areas. Marks et al. (1978) and Dadi and Marks (1987) discussed the detection of constant signals when Laplace noise is present. Hsu (1979) considered a Laplace distribution for position errors in navigation. Easterling (1978) considered a double exponential measurement error to study steam generator inspection. Other applications in communication theory, finance or environment sciences can be found in Kotz et al. (2001) and references therein. Most applications exploited a Laplace distribution due to its tails being heavier than those of a normal distribution. As mentioned in Gel (2010), it is necessary to consider broader alternative distributions than normal distributions and to include heavy-tailed alternative distributions. Hence, in the simulation, five location/scale families are considered: normal distributions, t-distributions with 10 degrees of freedom, Gumbel distributions, Cauchy distributions and logistic distributions. Of the five distributions, the Gumbel distribution is asymmetric and the remaining distributions are symmetric. To apply AD, CvM or Watson to every small data set, MLE of location and scale parameters is used as in Section 3.1.

Tables 3.13, 3.15 and 3.17 show the empirical power and size of the one-sided moment based tests and smoothing based tests when all small data sets come from the same distribution. When the alternative is Gumbel or Cauchy, both moment based tests and smoothing based tests have good power for all three edf-based gof tests. On the contrary, when all small data sets of size 5 are from the t-distribution, logistic or normal distributions, we notice that the moment based tests are biased. Especially, AD has the bias problem even if the sample size increases to 10. This phenomenon seems to be contrary to the usual expectation that AD is more powerful

than other edf-based gof tests.

Tables 3.14, 3.16 and 3.18 show the empirical power and size of the two-sided moment based tests when a significance level α_2=0.01 is used. From these tables, we notice that the bias problem is not completely solved. The effects of selecting the two significance levels are investigated through Figures 3.13 to 3.16. As in the previous section, the case of 100 data sets with 5 observations each is considered. We notice that, when tests are biased, the power of the two-sided moment based tests is higher than that of the one-sided moment based tests. Especially, the amount of relative increases in the power is the biggest when Fisher's method is applied to P-values from Watson. However, large relative increases in the power cannot guarantee higher power than the level of tests. For example, Fisher's method using P-values from Watson still has power less than 0.05 at most of the considered significance levels α_2. Also, we notice that the bias problem is not resolved even if evenly divided significance levels are used. Such an examination suggests that it may be better to use smoothing based tests rather than the two-sided moment based tests to handle the bias.

When smoothing based tests are applied to P-values from AD, there is an odd decrease in the power under the t-distribution as the sample size increases from 5 to 10. Such a decrease in the power does not exist, when CvM or Watson is used. This can be explained by the density of the P-value. Figures 3.17, 3.18 and 3.19 show the density of the P-value when data sets are from the t-distribution. When the sample size is 10, the density of the P-value from CvM or Watson has an increasing shape and shows a relatively bigger departure from uniformity. This is the reason that CvM or Watson has better power as the sample size increases. However, the density of the P-value from AD, when the sample size is 10, still does not show much difference from uniformity, indicating a possible decrease in the power.

We may expect that the power increases as the number of data sets increases because we have more information. This may not be true when tests are biased. For example, when Watson is applied to small data sets, with 5 observations, from logistic distributions, the power of moment based tests decreases as the number of data sets increases. This reflects an asymptotic failure of the one-sided moment based tests due to the bias. When smoothing based tests are used, however, the power increases as the number of data sets increases, indicating that these tests detect any departure from uniformity.

Under fixed alternatives, Edgington's method attains slightly better power than Fisher's method. The performance of smoothing based tests depends on alternative distributions and the type of edf-based gof tests. For example, when Watson is used, and data sets come from normal or logistic distributions, the smooth test is better than the order selection test, especially for sample sizes 5. However, when Watson is used, and data sets are from Gumbel distributions, the order selection test is better than the smooth test. If we compare the power of smoothing based tests and moment based tests, there is no clear winner. When tests are not biased, moment based tests perform better than smoothing based tests. On the other hand, smoothing based tests may be preferable when the tests are biased. Of the three edf-based gof test, CvM and Watson show better performance than AD.

Tables 3.19 to 3.24 show the local empirical power, i.e., the power when 90% of data sets are from Laplace distributions. Except for two distributions, Cauchy and Gumbel distributions, the local power of both moment based tests and smoothing based tests is just around the size of tests regardless of the type of gof tests used. Such results are interesting especially when we consider normal local alternatives. In Section 3.1, we notice that both moment based tests and smoothing based tests detect departures from normality well when either all or a few data sets are from

Laplace distributions. On the contrary, when the null is Laplace and 10% of data sets are from normal distributions, both moment based tests and smoothing based tests have the power around the level of tests 0.05, indicating that both tests cannot detect a departure from the null. Such a result implies that, when we consider two distributions and perform a gof test, different power can be obtained depending on which one of the two is considered as the null. The effects of the significance levels α_2 are investigated through Figures 3.20 to 3.24. When 10% of data come from the t-distribution, Gumbel distributions, or logistic distributions, there does not exist much difference between the power at significance levels α_2. When 10% of data come from a normal distribution and Fisher's method is applied to P-values from AD or CvM, the power is above the size of tests when the significance level α_2 is about 0.014.

Under local alternatives, the smooth test tends to attain better power than the order selection test. The performance of the two moment based tests depends on the distribution from which 10% of data sets come. For example, when 10% of small data sets with sample sizes 10 are from Gumbel distributions, Edgington's method has slightly higher power than Fisher's method. Fisher's method is more powerful than Edgington's method when 10% of data sets are from Cauchy distributions. Figure 3.25 indicates the reason of the reversal in performance of the two moment based tests. We notice that evidence against the null is much stronger when the alternative is Cauchy. In Section 2.3, we found that Edgington's method performs better when there exists slight or moderate evidence against the null. The empirical power and the density of the P-value agree with this finding.

In addition to considering the local alternative where 90% of data sets are from the null, the empirical power is obtained under local alternatives where other than 90% of data sets are from the null. Only two alternative distributions, Gumbel and

logistic distributions, are considered. These are selected because we might obtain different insights according to whether tests are biased. Note that tests are biased when the alternative is a logistic distribution, and tests are not biased when the alternative is a Gumbel distribution.

Figures 3.26, 3.27 and 3.28 show the empirical power at the significance level $\alpha = 0.05$ when the alternative is a mixture of Laplace and Gumbel distributions. When the sample size is 5, the results from CvM are similar to those from AD. The results are similar regardless of the type of gof tests when the sample size is 10. From Figure 3.26, we notice that when AD or CvM is applied to data sets with 5 observations and less than 50% of data sets are from the null, moment based tests are slightly better than smoothing based tests. However, when more than 50% of data sets are from the null, both moment based tests and smoothing based tests have power that is close to the size of tests. Interestingly, Figure 3.27 shows that Fisher's method attains the worst power when Watson is applied to data sets with 5 observations. When there are at least 300 data sets with sample sizes 10, from Figure 3.28, we notice that the power of smoothing based tests is close to that of moment based tests. Also, there is no clear winner between the two moment based tests or between the two smoothing based tests. However, when we have 100 data sets with 10 observations and 10% or 20% of data sets are from the null, Edgington's method seems to be the best.

When the alternative is a mixture of Laplace and logistic distributions and we have data sets with sample sizes 5, we need to consider the bias problem. Figures 3.29 and 3.30 show the empirical power when AD and Watson are applied to data sets with sample sizes 5. The result when CvM is used is not shown here because the empirical power is just around the size of tests for all combining methods. To deal with the bias, the two-sided moment based test at significance level α_2=0.01 is

applied. When AD is used, smoothing based tests dominate moment based tests. In contrast, when Watson is used, Fisher's method and the smooth test have similar power and dominate the remaining two. Edgington's method has the worst power. Figures 3.31 and 3.32 show the empirical power when AD or Watson is applied to data sets with sample sizes 10. Since only AD is biased under logistic alternatives, the two-sided moment based tests are used for Figure 3.31. From the figure, we notice that smoothing based tests are superior to moment based tests, and the smooth test is better among the two smoothing based tests. When Watson is used, the result is quite different. From Figure 3.32, we notice that Edgington's method is the best and the power of Fisher's method is just around the size of tests. Even if the result from CvM is not shown here, it is similar to that from Watson.

It is clear that, from the empirical power under local alternatives, the performance of tests depends on the number of data sets, the sample sizes, the alternatives, and the type of edf-based tests. It may almost be impossible to find the one best method. However, we notice that AD is biased under the logistic alternative when the sample size is 10, unlike CvM or Watson. This indicates that CvM or Watson may be preferable to AD. Also, under the logistic local alternatives, Watson seems to have more reliable power than CvM. Even if moment based tests are more powerful than smoothing based tests when we do not have the bias issue, smoothing based tests are better when we have the bias problem. Also, the power of smoothing based tests is just a little bit inferior to moment based tests when tests are not biased. Hence, when testing whether data come from Laplace distributions, applying smoothing based tests to P-values from Watson might be preferable.

Table 3.13: This table shows the size(%) and the power(%) of the test. The null hypothesis is that data come from Laplace distributions and AD is applied to every small data set. For moment based tests, the one-sided test is used.

		Edgington's method						Fisher's method					
n	p	Laplace	t(10)	Gumbel	Cauchy	Logistic	Normal	Laplace	t(10)	Gumbel	Cauchy	Logistic	Normal
5	100	5.3	2.2	14.6	98.9	1.8	2.3	5.9	2.2	14.8	100.0	2.5	3.3
	300	5.4	1.0	29.8	100.0	1.4	1.0	4.8	1.6	26.7	100.0	1.9	2.2
	500	4.6	0.8	39.8	100.0	0.9	1.2	4.7	1.4	38.2	100.0	1.5	2.0
	1000	4.8	0.1	61.8	100.0	0.2	0.3	4.6	0.4	60.1	100.0	0.5	0.8
10	100	4.8	3.4	77.1	100.0	1.9	6.2	5.0	2.3	67.2	100.0	1.3	4.2
	300	4.6	2.4	99.5	100.0	1.4	7.8	4.3	0.9	98.1	100.0	0.5	3.1
	500	4.8	2.2	100.0	100.0	1.4	8.8	4.0	0.4	99.9	100.0	0.4	3.2
	1000	5.4	1.8	100.0	100.0	0.2	11.5	4.4	0.3	100.0	100.0	0.0	2.1
		Smooth Test						Order Selection Test					
n	p	Laplace	t(10)	Gumbel	Cauchy	Logistic	Normal	Laplace	t(10)	Gumbel	Cauchy	Logistic	Normal
5	100	5.1	4.3	6.9	99.1	5.4	5.8	5.1	5.6	8.6	97.8	6.1	6.4
	300	5.8	7.7	16.9	100.0	8.1	8.8	5.3	8.8	17.4	100.0	8.7	8.5
	500	5.2	11.1	27.1	100.0	11.2	11.2	5.0	11.5	26.7	100.0	11.3	12.2
	1000	4.8	19.4	48.1	100.0	19.1	16.2	4.3	19.4	46.2	100.0	19.1	16.2
10	100	4.4	6.3	52.3	100.0	5.8	6.3	5.8	5.2	62.8	100.0	5.1	5.2
	300	5.2	8.0	98.2	100.0	8.8	8.6	5.0	6.3	98.2	100.0	7.9	6.8
	500	5.1	11.6	100.0	100.0	12.2	12.6	4.8	9.0	100.0	100.0	11.4	10.5
	1000	5.2	17.2	100.0	100.0	22.4	18.4	5.5	14.1	100.0	100.0	20.6	15.8

Table 3.14: This table shows the size(%) and the power(%) of the two-sided moment based test. The null hypothesis is that data come from Laplace distributions and AD is applied to every small data set. The significance level α_2=0.01 is used.

		Edgington's method						Fisher's method					
n	p	Laplace	t(10)	Gumbel	Cauchy	Logistic	Normal	Laplace	t(10)	Gumbel	Cauchy	Logistic	Normal
5	100	5.1	3.5	12.9	98.7	3.3	4.0	6.3	4.0	12.6	100.0	3.9	4.8
	300	5.5	4.8	26.4	100.0	5.6	4.8	5.1	4.0	23.2	100.0	5.1	5.0
	500	4.7	7.0	36.1	100.0	6.4	6.3	5.0	6.5	34.6	100.0	6.2	5.9
	1000	4.5	11.6	57.6	100.0	11.4	8.9	4.5	9.4	56.4	100.0	9.8	7.0
10	100	4.9	4.3	74.2	100.0	3.0	5.6	5.5	3.7	63.2	100.0	3.4	4.5
	300	4.4	3.8	99.2	100.0	4.0	6.6	4.2	4.0	97.4	100.0	5.5	4.0
	500	4.9	3.8	100.0	100.0	4.6	7.9	4.2	4.9	99.9	100.0	8.2	3.5
	1000	5.8	4.2	100.0	100.0	8.2	10.1	4.8	8.9	100.0	100.0	17.5	3.6

Table 3.15: This table shows the size(%) and the power(%) of the test. The null hypothesis is that data come from Laplace distributions and CvM is applied to every small data set. For moment based tests, the one-sided test is used.

		Edgington's method						Fisher's method					
n	p	Laplace	t(10)	Gumbel	Cauchy	Logistic	Normal	Laplace	t(10)	Gumbel	Cauchy	Logistic	Normal
5	100	5.4	3.8	17.4	96.9	3.4	5.1	6.0	3.6	15.8	99.9	3.1	5.1
	300	5.1	3.6	35.9	100.0	3.4	4.6	4.5	3.2	30.1	100.0	3.6	5.1
	500	5.4	3.2	48.8	100.0	3.4	5.9	5.0	2.9	42.6	100.0	3.0	4.6
	1000	5.6	2.9	72.6	100.0	1.7	5.5	4.5	2.4	66.8	100.0	1.3	4.5
10	100	5.3	11.2	74.8	100.0	6.9	23.6	5.3	6.6	64.2	100.0	3.8	14.4
	300	4.4	18.1	99.0	100.0	12.2	47.1	4.0	7.4	97.2	100.0	5.1	23.6
	500	4.8	22.1	100.0	100.0	13.2	66.8	3.6	8.7	99.9	100.0	5.3	35.3
	1000	4.6	39.5	100.0	100.0	17.2	88.5	4.3	13.6	100.0	100.0	4.8	52.8
		Smooth Test						Order Selection Test					
n	p	Laplace	t(10)	Gumbel	Cauchy	Logistic	Normal	Laplace	t(10)	Gumbel	Cauchy	Logistic	Normal
5	100	4.8	3.8	8.4	96.8	5.2	5.2	4.3	4.3	10.2	94.8	5.0	5.5
	300	5.3	4.8	20.1	100.0	5.2	5.1	5.0	5.0	22.3	100.0	5.7	4.9
	500	5.4	5.7	32.6	100.0	4.8	5.6	5.1	5.3	33.0	100.0	5.3	5.8
	1000	4.6	6.2	60.6	100.0	7.0	5.5	5.1	5.8	57.6	100.0	6.8	5.8
10	100	5.1	6.2	49.4	100.0	5.5	10.3	5.4	6.6	59.4	100.0	5.2	15.2
	300	5.0	10.3	97.4	100.0	8.8	31.0	5.1	11.6	97.5	100.0	8.8	34.8
	500	5.2	15.4	100.0	100.0	10.6	51.5	5.1	15.6	100.0	100.0	10.5	55.0
	1000	4.8	28.9	100.0	100.0	14.4	80.2	5.1	31.2	100.0	100.0	14.8	81.8

Table 3.16: This table shows the size(%) and the power(%) of the two-sided moment based test. The null hypothesis is that data come from Laplace distributions and CvM is applied to every small data set. The significance level α_2=0.01 is used.

		Edgington's method						Fisher's method					
n	p	Laplace	t(10)	Gumbel	Cauchy	Logistic	Normal	Laplace	t(10)	Gumbel	Cauchy	Logistic	Normal
5	100	5.1	3.8	14.9	96.4	4.2	4.6	6.4	3.8	14.1	99.8	3.5	5.5
	300	5.2	3.9	32.1	100.0	4.0	4.4	4.8	4.0	26.9	100.0	4.3	5.2
	500	5.0	4.0	44.1	100.0	4.0	5.4	5.1	4.5	39.1	100.0	4.0	4.2
	1000	5.7	3.5	69.1	100.0	3.5	5.1	4.8	4.0	62.5	100.0	3.8	4.3
10	100	4.8	9.6	71.0	100.0	5.9	20.5	5.2	5.6	59.8	100.0	3.8	11.6
	300	4.0	15.3	98.8	100.0	10.6	43.5	3.8	6.2	96.3	100.0	4.7	20.0
	500	4.8	19.5	100.0	100.0	11.2	63.2	3.4	7.3	99.9	100.0	4.6	30.7
	1000	5.0	34.9	100.0	100.0	14.5	85.7	5.0	11.8	100.0	100.0	4.4	48.9

Table 3.17: This table shows the size(%) and the power(%) of the test. The null hypothesis is that data come from Laplace distributions and Watson is applied to every small data set. For moment based tests, the one-sided test is used.

		Edgington's method						Fisher's method					
n	p	Laplace	t(10)	Gumbel	Cauchy	Logistic	Normal	Laplace	t(10)	Gumbel	Cauchy	Logistic	Normal
5	100	4.7	3.1	12.6	98.3	2.5	4.0	5.1	1.4	5.4	100.0	0.9	0.8
	300	4.8	2.4	25.1	100.0	2.2	2.6	5.0	0.2	10.2	100.0	0.7	0.4
	500	5.4	1.6	32.9	100.0	2.1	3.9	5.0	0.2	10.5	100.0	0.2	0.1
	1000	5.8	1.2	53.8	100.0	0.8	2.6	5.5	0.0	15.2	100.0	0.0	0.1
10	100	5.1	15.1	71.0	100.0	8.9	33.2	4.8	7.8	54.8	100.0	5.0	21.6
	300	4.9	27.0	98.8	100.0	16.3	66.2	4.7	12.8	93.0	100.0	6.6	42.7
	500	4.6	35.6	100.0	100.0	19.1	84.2	4.5	15.4	99.4	100.0	7.6	62.1
	1000	4.2	59.3	100.0	100.0	28.7	97.6	4.2	25.9	100.0	100.0	7.5	85.0
		Smooth Test						Order Selection Test					
n	p	Laplace	t(10)	Gumbel	Cauchy	Logistic	Normal	Laplace	t(10)	Gumbel	Cauchy	Logistic	Normal
5	100	5.2	5.9	6.8	100.0	6.2	6.6	4.3	5.2	8.1	98.8	5.3	6.1
	300	5.1	11.6	16.6	100.0	9.8	13.8	5.1	7.2	18.6	100.0	7.8	8.8
	500	5.2	14.8	22.9	100.0	12.6	19.8	5.0	9.7	24.6	100.0	9.8	13.3
	1000	5.2	24.8	43.2	100.0	21.1	36.3	5.4	18.4	45.8	100.0	16.4	27.2
10	100	5.1	7.6	45.4	100.0	5.6	14.1	5.4	8.9	56.5	100.0	5.7	20.8
	300	5.1	15.3	95.7	100.0	8.7	45.9	4.8	16.6	96.3	100.0	9.8	50.7
	500	5.0	21.8	99.8	100.0	12.2	70.5	5.1	23.8	99.7	100.0	13.6	72.0
	1000	5.6	43.6	100.0	100.0	19.5	94.5	5.5	44.6	100.0	100.0	20.7	94.3

Table 3.18: This table shows the size(%) and the power(%) of the two-sided moment based test. The null hypothesis is that data come from Laplace distributions and Watson is applied to every small data set. The significance level α_2=0.01 is used.

		Edgington's method						Fisher's method					
n	p	Laplace	t(10)	Gumbel	Cauchy	Logistic	Normal	Laplace	t(10)	Gumbel	Cauchy	Logistic	Normal
5	100	4.6	3.2	10.4	97.9	3.5	4.4	5.5	3.5	5.1	100.0	3.5	2.9
	300	5.2	4.2	22.2	100.0	4.0	3.2	5.5	6.3	8.6	100.0	6.6	5.7
	500	5.1	3.2	29.0	100.0	4.2	4.2	4.8	10.4	9.3	100.0	10.3	10.5
	1000	6.0	4.7	48.8	100.0	4.6	3.0	5.4	21.9	13.1	100.0	20.8	22.2
10	100	4.8	12.7	67.2	100.0	7.5	30.0	4.7	6.5	50.0	100.0	3.9	18.6
	300	5.2	23.5	98.2	100.0	14.2	62.0	5.0	10.7	92.0	100.0	6.0	37.9
	500	4.6	31.4	100.0	100.0	16.2	81.5	5.0	13.2	99.1	100.0	6.6	58.3
	1000	4.7	54.6	100.0	100.0	25.3	96.8	5.1	22.8	100.0	100.0	6.9	82.7

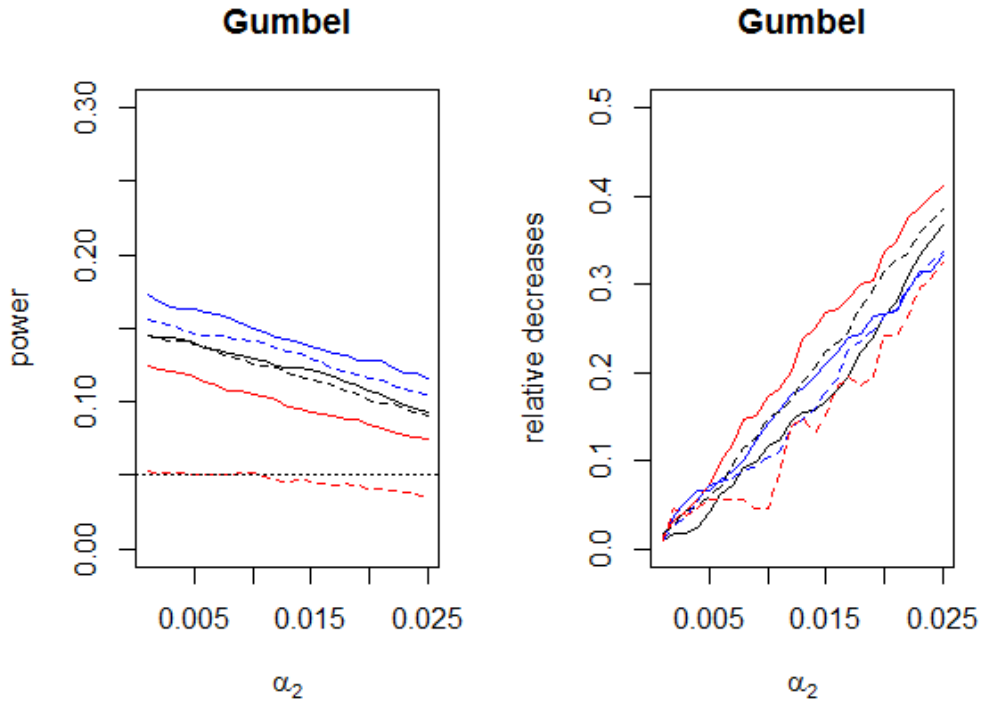

Figure 3.13: The left and right plots show the power of the two-sided tests and the relative power decrease over various significance levels α_2, when there are 100 data sets with 5 observations. In both plots, black, blue and red lines represent AD, CvM, and Watson, respectively. Also, solid and dashed lines represent Edgington's method and Fisher's method, respectively. The dotted line in the left plot denotes the significance level, 0.05.

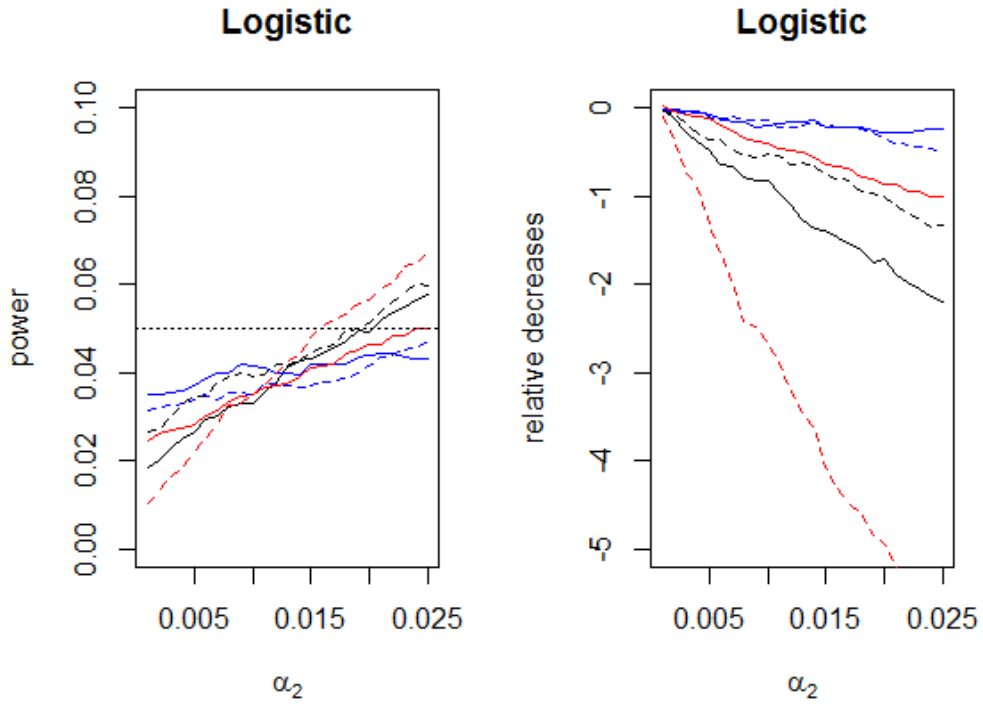

Figure 3.14: The left and right plots show the power of the two-sided moment based tests and the relative power decrease over various significance levels α_2, when there are 100 data sets with 5 observations. In both plots, black, blue and red lines represent AD, CvM, and Watson, respectively. Also, the solid and dashed lines represent Edgington's method and Fisher's method, respectively. The dotted line in the left plot denotes the significance level, 0.05.

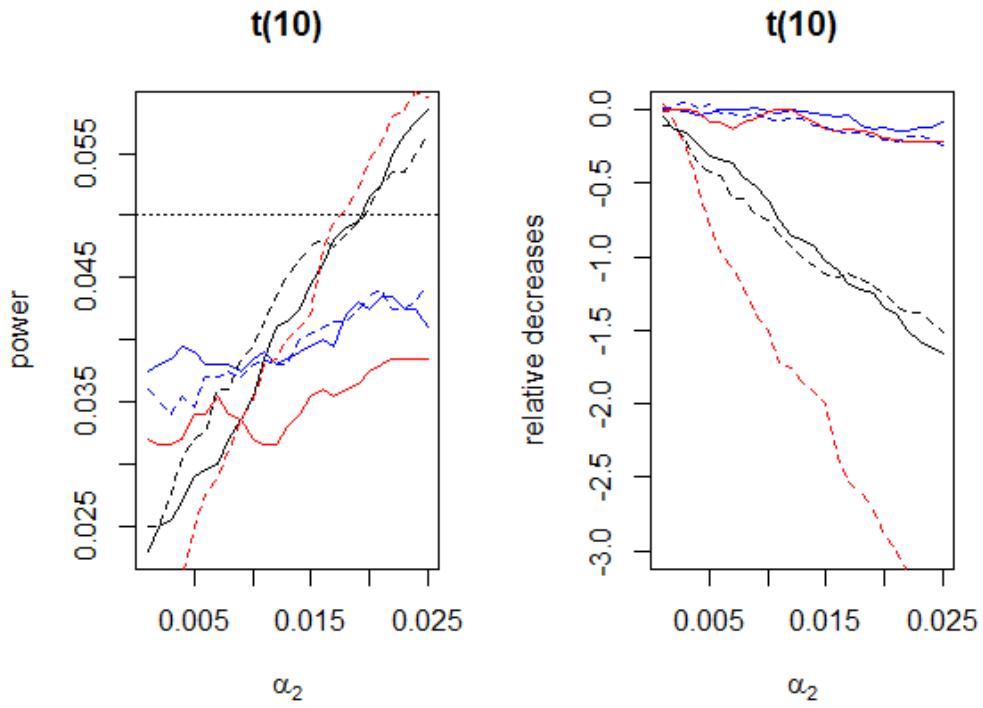

Figure 3.15: The left and right plots show the power of the two-sided moment based tests and the relative power decrease over various significance levels α_2, when there are 100 data sets with 5 observations. In both plots, black, blue and red lines represent AD, CvM, and Watson, respectively. Also, the solid and dashed lines represent Edgington's method and Fisher's method, respectively. The dotted line in the left plot denotes the significance level, 0.05.

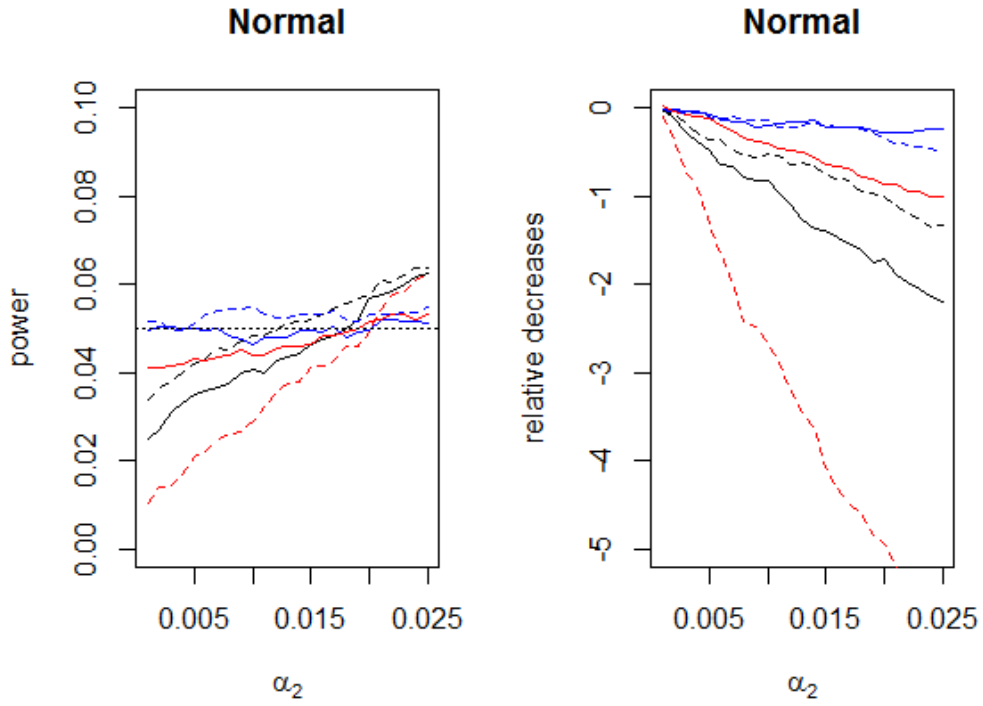

Figure 3.16: The left and right plots show the power of the two-sided moment tests and the relative power decrease over various significance levels α_2, when there are 100 data sets with 5 observations. In both plots, black, blue and red lines represent AD, CvM, and Watson, respectively. Also, the solid and dashed lines represent Edgington's method and Fisher's method, respectively. The dotted line in the left plot denotes the significance level, 0.05.

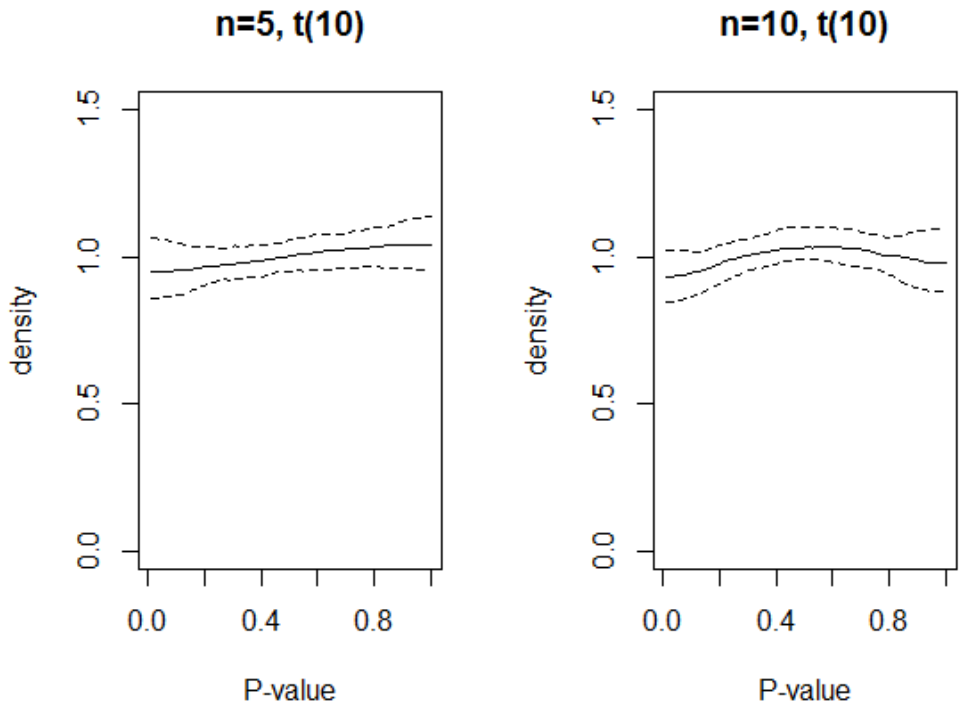

Figure 3.17: This figure shows the density of the P-value when AD is applied and the alternative distribution is the t-distribution with 10 degrees of freedom. The solid line is the median of kernel density estimates and the dashed lines are 0.025 and 0.975 percentiles of kernel density estimates.

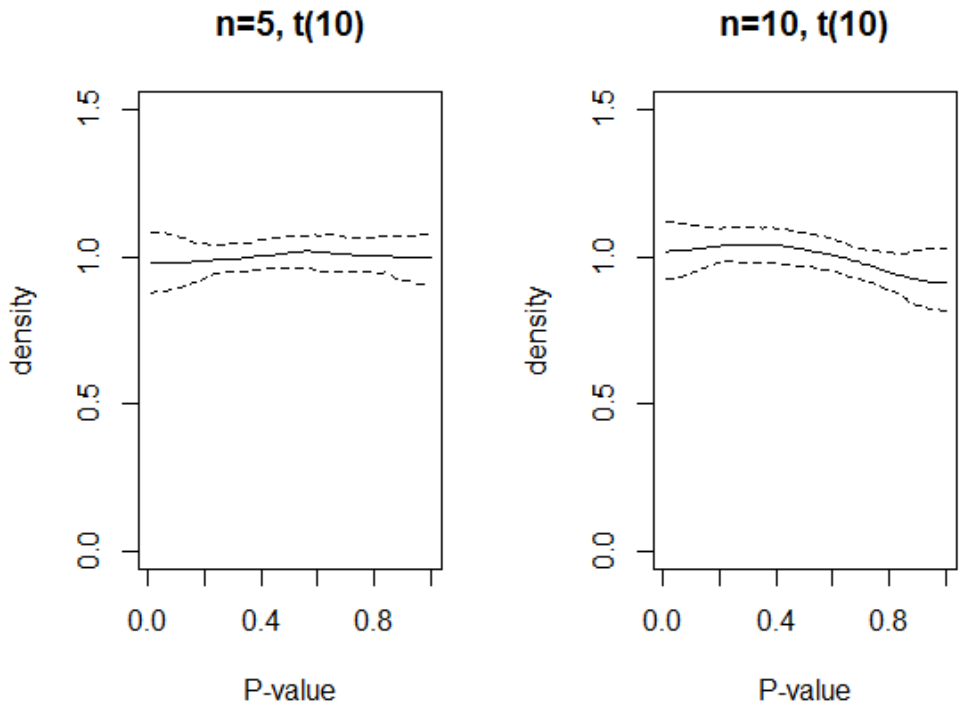

Figure 3.18: This figure shows the density of the P-value when CvM is applied and the alternative distribution is the t-distribution with 10 degrees of freedom. The solid line is the median of kernel density estimates and the dashed lines are 0.025 and 0.975 percentiles of kernel density estimates.

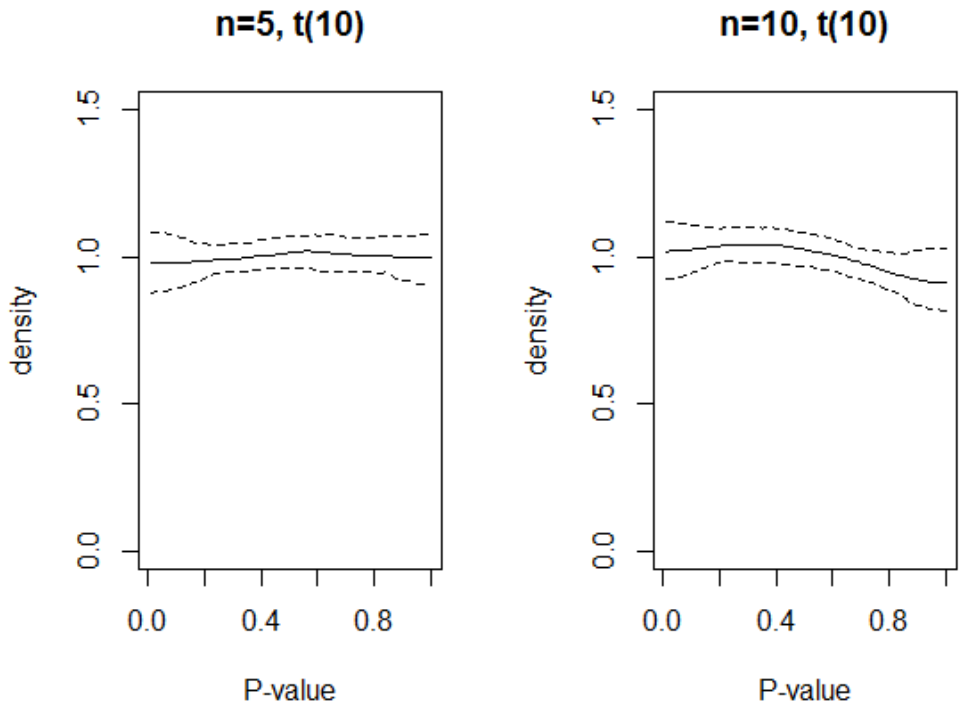

Figure 3.19: This figure shows the density of the P-value when Watson is applied and the alternative distribution is the t-distribution with 10 degrees of freedom. The solid line is the median of kernel density estimates and the dashed lines are 0.025 and 0.975 percentiles of kernel density estimates.

Table 3.19: This table shows the local power of the test when 90% of data sets are from the null distribution. The null hypothesis is that data come from Laplace distributions and AD is applied to every small data set. For moment based tests, the one-sided test is used.

		Edgington's method					Fisher's method				
n	p	t(10)	Gumbel	Cauchy	Logistic	Normal	t(10)	Gumbel	Cauchy	Logistic	Normal
5	100	5.2	4.8	11.3	4.4	4.2	4.2	5.1	17.2	4.8	4.1
	300	4.2	5.9	16.4	4.7	5.2	4.8	6.3	31.3	4.4	5.1
	500	4.0	5.8	23.2	4.0	3.7	3.7	6.2	43.6	3.6	3.8
	1000	3.3	6.4	35.2	3.6	3.6	3.6	6.2	65.5	3.8	3.8
10	100	4.5	7.3	19.9	4.4	6.0	4.2	7.3	58.5	4.8	5.1
	300	3.8	10.3	39.8	4.2	4.4	2.8	9.6	91.3	3.4	3.1
	500	4.6	13.6	53.7	4.0	5.5	3.6	11.2	98.0	3.1	4.0
	1000	4.3	17.8	80.0	4.2	4.6	2.9	14.2	100.0	2.7	3.4

		Smooth Test					Order Selection Test				
n	p	t(10)	Gumbel	Cauchy	Logistic	Normal	t(10)	Gumbel	Cauchy	Logistic	Normal
5	100	5.50	4.05	7.75	4.70	5.30	5.20	4.60	7.45	4.75	5.05
	300	4.95	4.85	11.05	5.35	4.95	4.65	4.70	9.30	4.95	4.70
	500	4.65	4.80	16.85	5.50	5.65	4.75	4.60	13.65	5.65	5.60
	1000	5.35	5.50	25.00	5.35	5.20	5.50	5.20	21.90	5.10	4.45
10	100	4.85	4.90	19.90	5.00	5.40	4.75	5.15	11.75	4.75	5.25
	300	5.40	6.50	40.90	5.65	5.05	5.20	6.65	26.05	5.60	5.05
	500	5.00	7.65	56.80	5.00	5.25	5.10	7.75	37.80	4.75	5.50
	1000	6.15	10.30	86.70	5.10	4.70	6.00	9.95	69.60	5.30	4.60

Table 3.20: This table shows the local power(%) of the two-sided moment based test when 90% of data sets are from the null distribution. The null hypothesis is that data come from Laplace distributions and AD is applied to every small data set. The significance level α_2=0.01 is used.

		Edgington's method					Fisher's method				
n	p	t(10)	Gumbel	Cauchy	Logistic	Normal	t(10)	Gumbel	Cauchy	Logistic	Normal
5	100	5.1	4.2	10.2	4.7	4.8	4.2	4.6	15.4	4.8	4.9
	300	4.1	5.5	13.3	4.6	5.1	5.2	5.9	27.3	4.8	4.6
	500	4.5	4.6	20.2	5.1	3.8	4.3	5.2	39.9	4.5	4.5
	1000	4.3	5.6	31.4	4.4	3.9	4.8	5.3	61.8	4.5	4.0
10	100	5.1	6.6	17.1	4.3	5.5	4.8	6.6	55.0	5.1	4.8
	300	4.7	9.0	35.7	4.8	4.5	3.8	7.8	89.3	4.1	4.0
	500	4.3	12.0	49.1	4.3	5.0	4.6	9.3	97.5	4.0	4.2
	1000	5.2	15.2	77.3	4.5	4.3	4.3	11.6	100.0	4.0	4.1

Table 3.21: This table shows the local power of the test when 90% of data sets are from the null distribution. The null hypothesis is that data come from Laplace distributions and CvM is applied to every small data set. For moment based tests, the one-sided test is used.

		Edgington's method					Fisher's method				
n	p	t(10)	Gumbel	Cauchy	Logistic	Normal	t(10)	Gumbel	Cauchy	Logistic	Normal
5	100	5.3	5.0	10.4	5.1	4.2	4.7	5.0	15.5	4.5	4.4
	300	4.3	6.0	15.0	4.7	5.7	5.2	6.2	26.1	4.8	5.1
	500	4.6	6.2	21.2	5.1	5.1	4.2	6.2	36.2	4.0	4.5
	1000	4.9	7.1	31.9	5.0	5.2	4.0	7.2	55.5	4.8	4.6
10	100	5.6	7.6	17.7	5.1	6.0	4.9	7.1	55.4	4.8	5.6
	300	5.3	10.3	34.2	5.2	6.3	3.7	8.8	88.2	4.0	4.6
	500	5.5	12.6	46.9	5.4	6.6	4.7	10.8	97.4	3.8	5.6
	1000	6.5	16.9	72.5	4.8	8.0	4.2	13.5	99.9	3.6	5.4

		Smooth Test					Order Selection Test				
n	p	t(10)	Gumbel	Cauchy	Logistic	Normal	t(10)	Gumbel	Cauchy	Logistic	Normal
5	100	5.7	5.3	7.8	4.9	5.1	5.3	4.4	6.6	4.5	4.8
	300	4.8	5.0	10.1	4.7	5.1	4.3	4.5	8.6	4.5	4.8
	500	4.8	4.4	14.0	5.9	5.1	5.1	4.3	12.3	5.9	4.8
	1000	4.2	5.1	22.1	5.2	5.0	4.6	5.1	19.6	5.0	4.8
10	100	4.8	4.4	19.1	4.3	4.2	4.6	5.6	10.3	4.8	5.7
	300	5.7	5.9	35.9	5.3	5.2	5.6	6.2	20.8	5.3	5.4
	500	5.0	7.3	50.0	5.0	5.5	4.7	7.2	30.8	4.9	5.7
	1000	6.2	9.6	81.5	5.1	5.4	6.8	9.8	61.9	4.4	6.0

Table 3.22: This table shows the local power(%) of the two-sided moment based test when 90% of data sets are from the null distribution. The null hypothesis is that data come from Laplace distributions and CvM is applied to every small data set. The significance level α_2=0.01 is used.

		Edgington's method					Fisher's method				
n	p	t(10)	Gumbel	Cauchy	Logistic	Normal	t(10)	Gumbel	Cauchy	Logistic	Normal
5	100	5.3	4.3	9.0	5.0	4.6	4.3	4.4	14.1	4.8	4.7
	300	4.2	5.5	12.3	4.8	5.4	5.1	6.0	22.4	4.8	5.1
	500	4.7	5.4	18.9	5.3	4.9	4.7	5.6	33.2	4.6	4.0
	1000	4.3	6.4	28.1	5.0	5.3	4.8	6.1	50.8	4.8	4.6
10	100	5.4	7.2	15.2	4.8	5.7	5.3	6.8	52.0	5.1	5.4
	300	5.2	8.6	30.6	5.4	5.4	4.4	7.2	86.2	4.2	4.8
	500	4.7	10.5	42.1	5.2	6.0	4.8	9.3	96.9	4.2	5.4
	1000	6.3	14.3	68.7	4.8	6.7	5.2	11.0	99.9	4.0	4.7

Table 3.23: This table shows the local power of the test when 90% of data sets are from the null distribution. The null hypothesis is that data come from Laplace distributions and Watson is applied to every small data set. For moment based tests, the one-sided test is used.

		Edgington's method					Fisher's method				
n	p	t(10)	Gumbel	Cauchy	Logistic	Normal	t(10)	Gumbel	Cauchy	Logistic	Normal
5	100	4.9	5.2	11.4	5.2	4.5	4.2	5.1	27.6	4.6	3.5
	300	4.3	5.6	17.2	4.4	5.1	4.3	5.4	51.6	3.4	3.8
	500	3.8	5.8	25.1	5.1	5.0	3.3	4.8	70.0	3.8	3.1
	1000	4.6	6.9	36.3	4.5	5.4	3.5	5.7	90.5	3.6	4.2
10	100	6.0	7.6	18.9	5.5	6.4	4.7	6.6	70.7	5.0	5.3
	300	5.8	10.2	37.5	5.8	6.3	4.0	7.7	96.8	4.4	5.4
	500	6.6	11.3	50.5	5.7	7.0	5.0	9.6	99.6	4.2	6.3
	1000	7.0	14.9	76.0	5.2	8.5	4.5	10.6	100.0	4.0	5.9

		Smooth Test					Order Selection Test				
n	p	t(10)	Gumbel	Cauchy	Logistic	Normal	t(10)	Gumbel	Cauchy	Logistic	Normal
5	100	4.8	5.2	10.0	5.5	5.2	5.5	5.1	6.8	5.4	5.0
	300	4.8	4.3	15.4	5.5	4.8	4.3	4.8	10.4	5.2	4.4
	500	5.3	5.2	21.3	5.4	4.5	5.2	5.3	14.6	5.5	5.2
	1000	5.3	5.4	34.6	4.8	5.1	5.0	5.4	23.7	5.1	5.1
10	100	5.2	5.1	25.1	4.8	4.9	5.7	5.2	9.6	4.8	5.0
	300	5.2	6.1	52.5	4.6	5.3	4.6	6.1	24.8	5.2	5.0
	500	5.1	6.2	70.4	6.0	5.6	4.8	6.8	36.0	5.5	5.3
	1000	5.3	8.5	96.6	5.0	6.1	4.8	8.8	72.4	4.9	5.8

Table 3.24: This table shows the local power(%) of the two-sided moment based test when 90% of data sets are from the null distributions. The null hypothesis is that data come from Laplace distributions and Watson is applied to every small data set. The significance level α_2=0.01 is used.

		Edgington's method					Fisher's method				
n	p	t(10)	Gumbel	Cauchy	Logistic	Normal	t(10)	Gumbel	Cauchy	Logistic	Normal
5	100	4.8	5.4	9.2	5.0	4.7	4.8	5.1	24.4	4.9	4.1
	300	4.3	5.0	15.2	4.3	5.0	4.5	5.2	48.3	4.1	4.5
	500	4.2	5.4	22.1	5.2	5.1	3.9	4.5	66.1	4.0	3.6
	1000	4.6	6.3	32.7	4.6	5.3	4.4	5.3	88.6	4.3	4.7
10	100	5.5	7.0	15.8	5.3	5.7	5.4	6.1	68.0	4.9	4.5
	300	5.6	9.0	33.9	5.8	5.7	4.2	6.6	96.4	4.8	5.5
	500	5.9	10.0	45.3	5.5	6.6	5.1	8.6	99.6	4.6	5.4
	1000	6.0	12.6	72.4	4.6	7.3	5.1	9.0	100.0	4.0	5.3

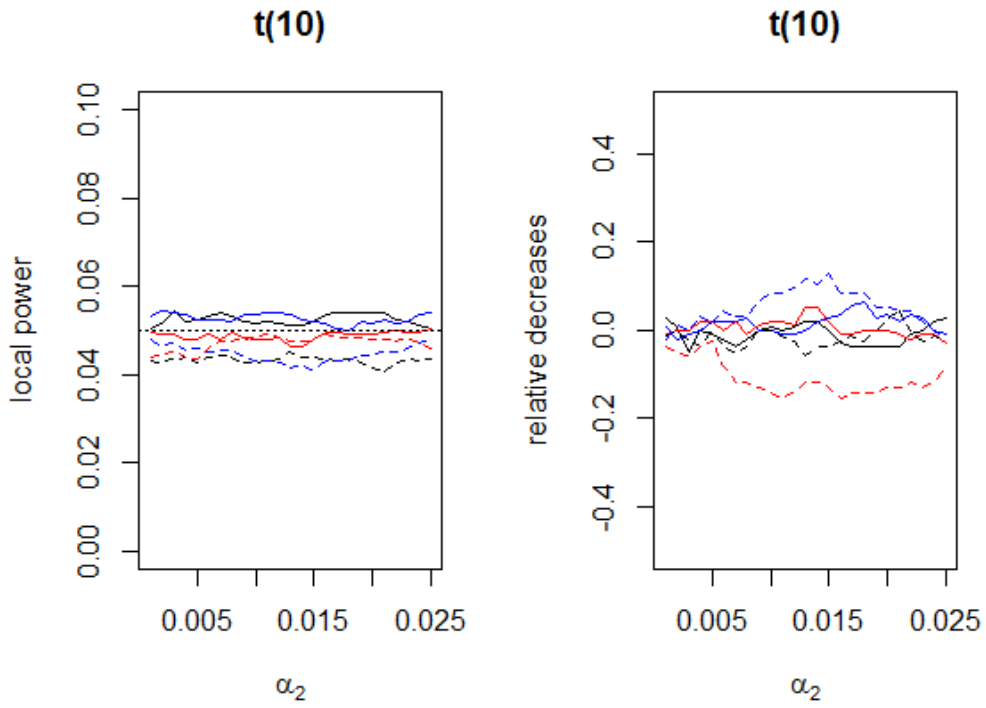

Figure 3.20: The left and right plots show the local power of the two-sided moment based tests and the relative power decrease over various significance levels α_2, when there are 100 data sets with 5 observations. In both plots, black, blue and red lines represent AD, CvM, and Watson, respectively. Also, the solid and dashed lines represent Edgington's method and Fisher's method, respectively. The dotted line in the left plot denotes the significance level, 0.05..

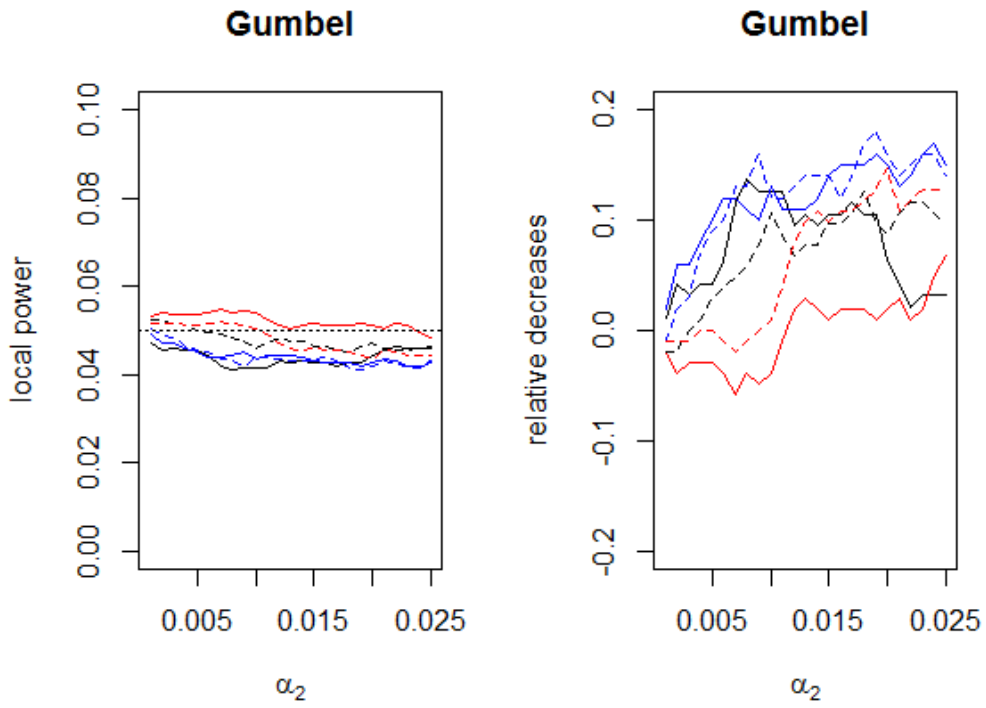

Figure 3.21: The left and right plots show the local power of the two-sided moment based tests and the relative power decrease over various significance levels α_2, when there are 100 data sets with 5 observations. In both plots, black, blue and red lines represent AD, CvM, and Watson, respectively. Also, the solid and dashed lines represent Edgington's method and Fisher's method, respectively. The dotted line in the left plot denotes the significance level, 0.05.

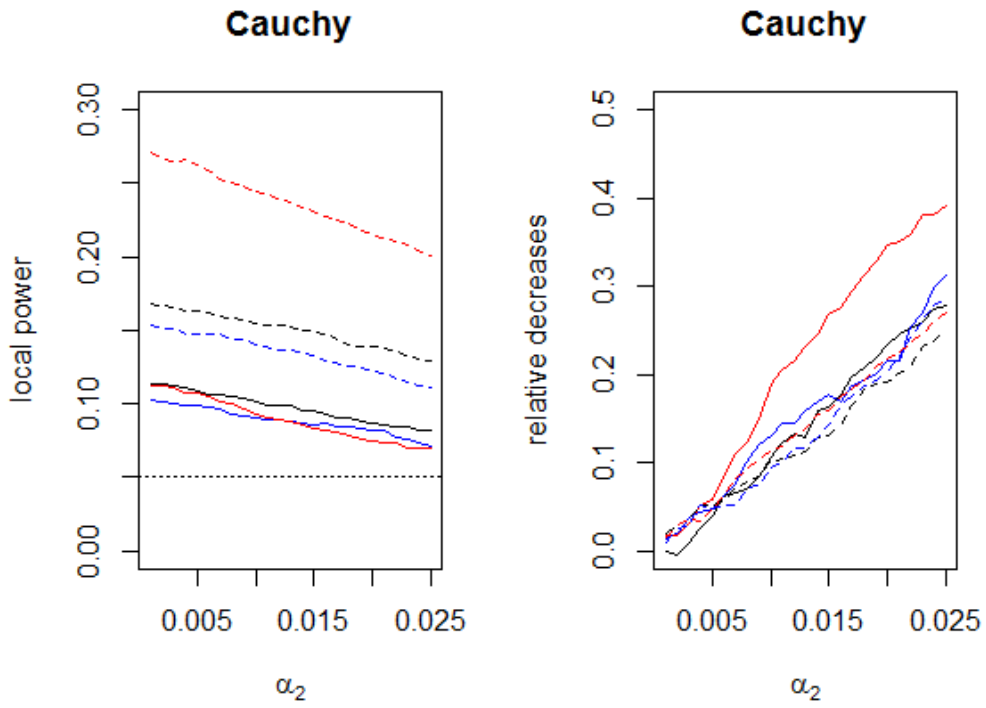

Figure 3.22: The left and right plots show the local power of the two-sided moment based tests and the relative power decrease over various significance levels, α_2, when there are 100 data sets with 5 observations. In both plots, black, blue and red lines represent AD, CvM, and Watson, respectively. Also, the solid and dashed lines represent Edgington's method and Fisher's method, respectively. The dotted line in the left plot denotes the significance level, 0.05.

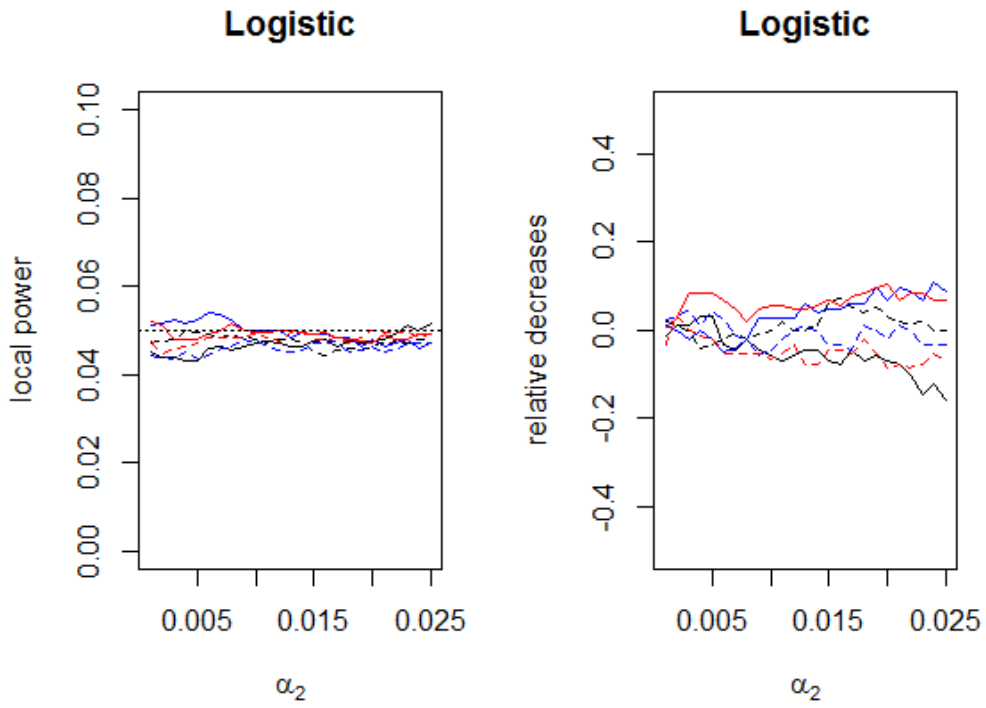

Figure 3.23: The left and right plots show the power of the two-sided moment based tests and the relative power decrease over various significance levels α_2, when there are 100 data sets with 5 observations. In both plots, black, blue and red lines represent AD, CvM, and Watson, respectively. Also, the solid and dashed lines represent Edgington's method and Fisher's method, respectively. The dotted line in the left plot denotes the significance level, 0.05.

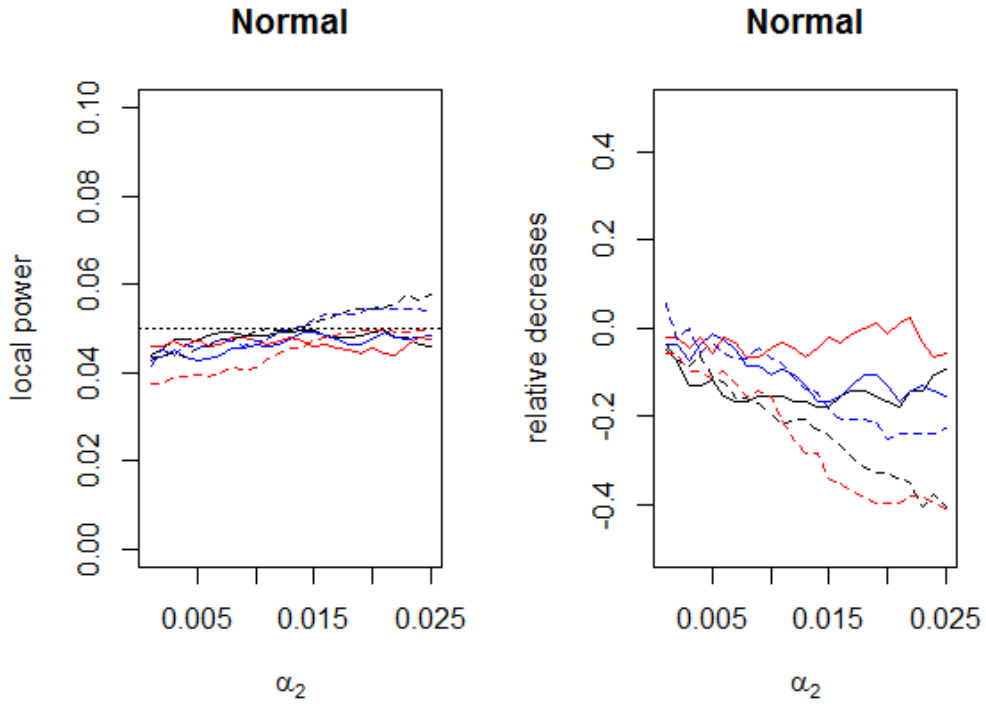

Figure 3.24: The left and right plots show the local power of the two-sided moment based tests and the relative power decrease over various significance levels α_2, when there are 100 data sets with 5 observations. In both plots, black, blue and red lines represent AD, CvM, and Watson, respectively. Also, the solid and dashed lines represent Edgington's method and Fisher's method, respectively. The dotted line in the left plot denotes the significance level, 0.05.

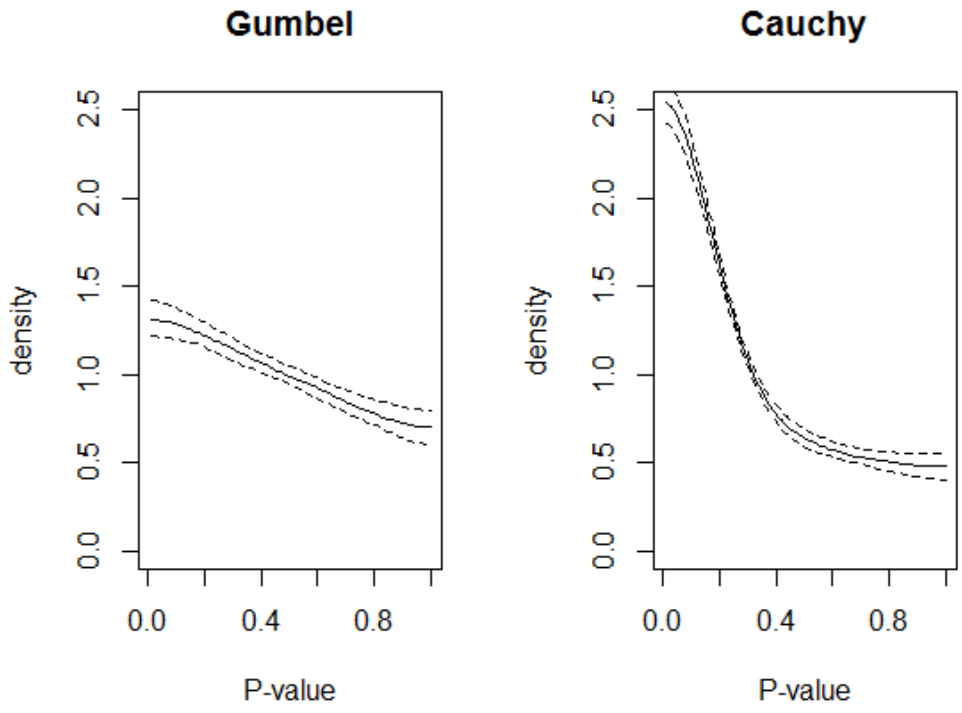

Figure 3.25: This figure shows the density of the P-value when CvM is applied to every small data set with sample sizes 10. The solid line is the median of kernel density estimates and the dashed lines are 0.025 percentiles and 0.975 percentiles of kernel density estimates.

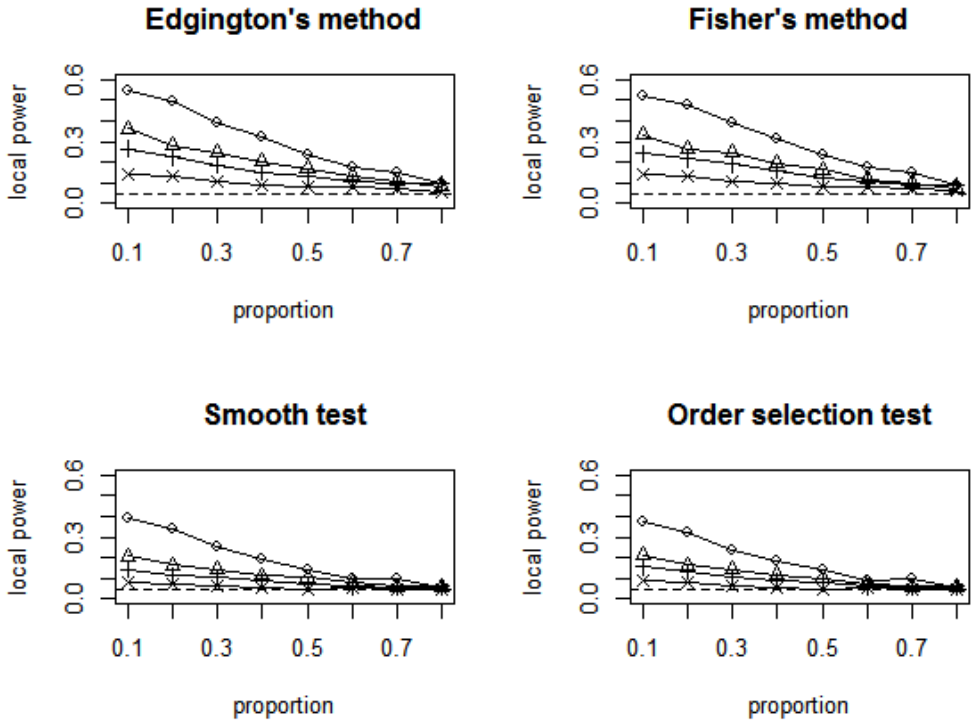

Figure 3.26: This figure shows the empirical power at the significance level 0.05 when testing whether data come from Laplace distributions, and the alternative is a mixture of Laplace and Gumbel distributions. The numbers of data sets considered are 100, 300, 500 and 1000, and the cross, plus, triangle and circle represent the number of data sets, respectively. AD is applied to every small data with sample sizes 5.

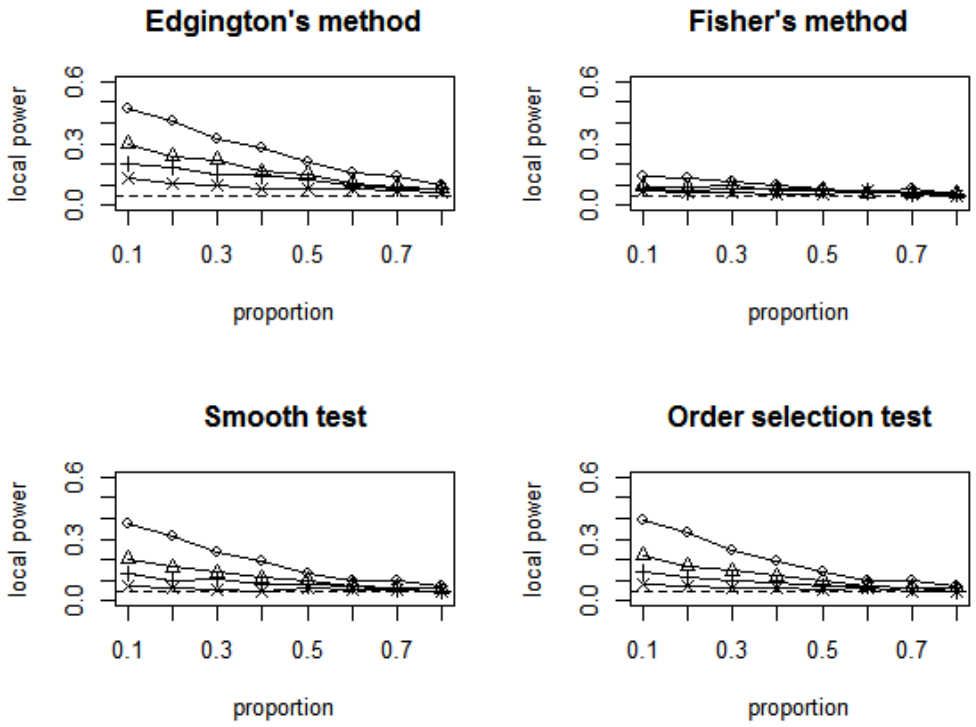

Figure 3.27: This figure shows the empirical power at the significance level 0.05 when testing whether data come from Laplace distributions, and the alternative is a mixture of Laplace and Gumbel distributions. The numbers of data sets considered are 100, 300, 500 and 1000, and the cross, plus, triangle and circle represent the number of data sets, respectively. Watson is applied to every small data with sample sizes 5.

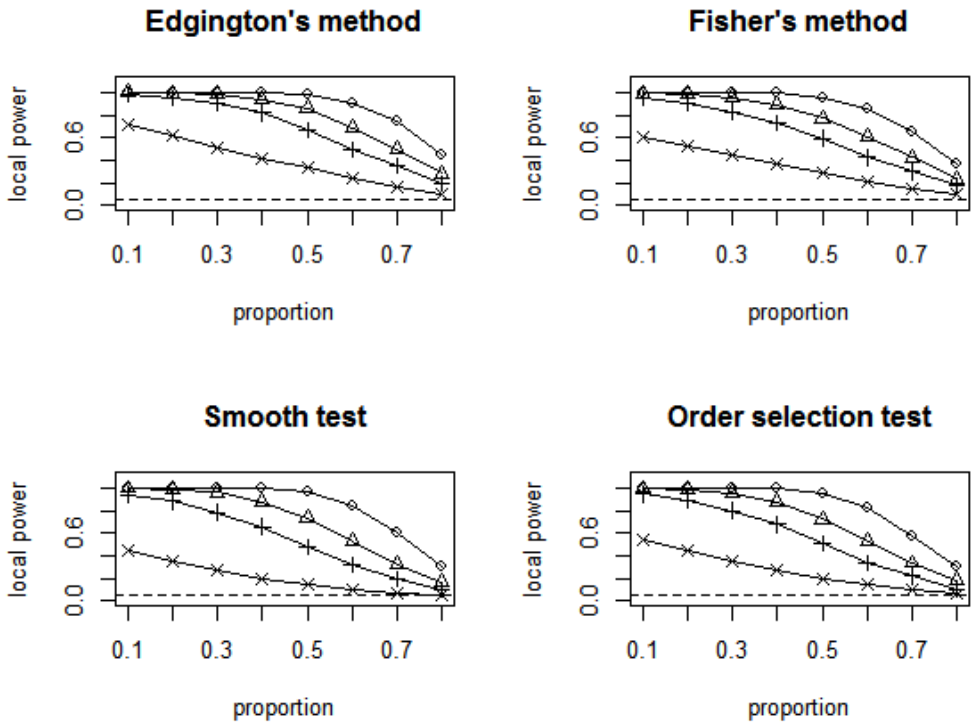

Figure 3.28: This figure shows the empirical power at the significance level 0.05 when testing whether data come from Laplace distributions, and the alternative is a mixture of Laplace and Gumbel distributions. The number of data sets considered are 100, 300, 500 and 1000, and the cross, plus, triangle and circle represent the number of data sets, respectively. AD is applied to every small data with sample sizes 10.

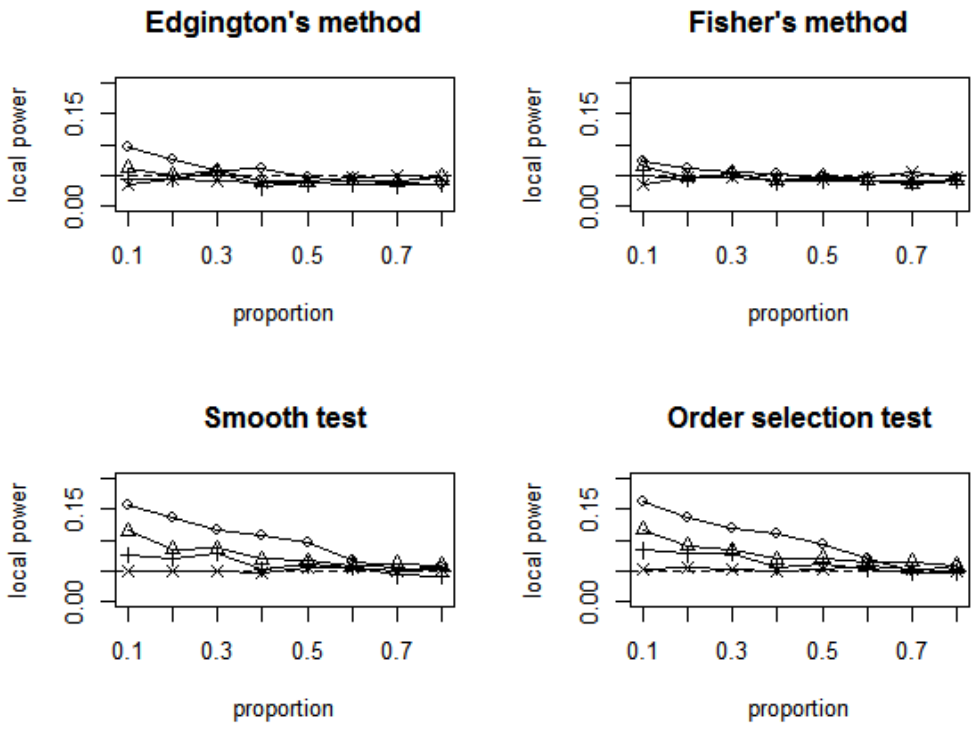

Figure 3.29: This figure shows the empirical power at the significance level 0.05 when testing whether data come from Laplace distributions, and the alternative is a mixture of Laplace and logistic distributions. The number of data sets considered are 100, 300, 500 and 1000, and the cross, plus, triangle and circle represent the number of data sets, respectively. For moment based tests, the two-sided test is used at the significance level α_2=0.01. AD is applied to every small data set with sample sizes 5.

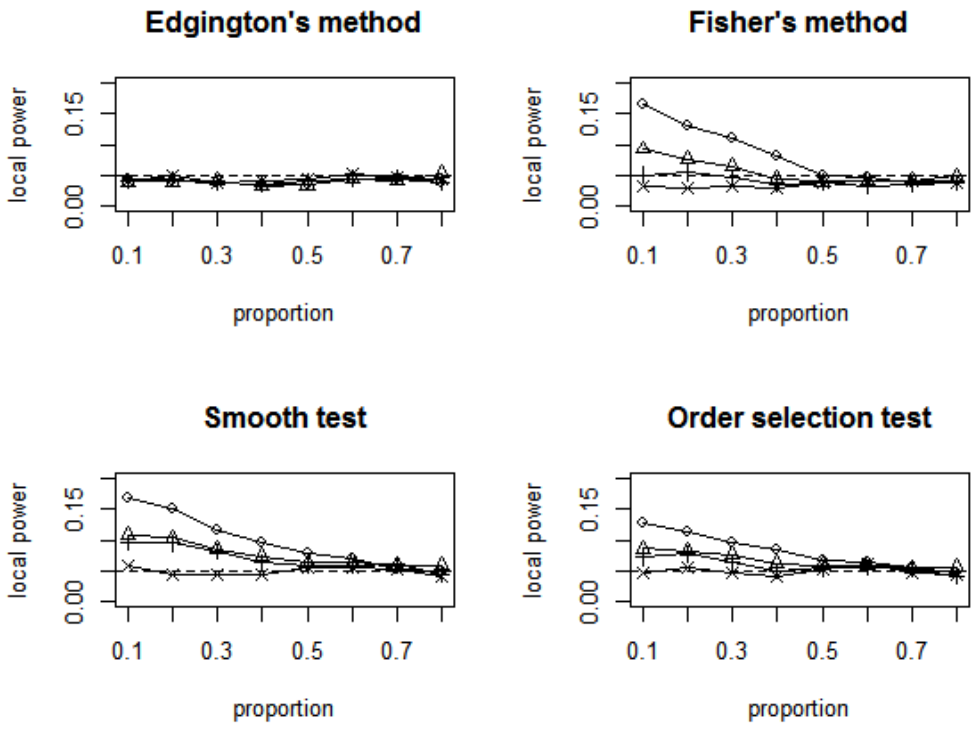

Figure 3.30: This figure shows the empirical power at the significance level 0.05 when testing whether data come from Laplace distributions, and the alternative is a mixture of Laplace and logistic distributions. The number of data sets considered are 100, 300, 500 and 1000, and the cross, plus, triangle and circle represent the number of data sets, respectively. For moment based tests, the two-sided test is used at the significance level α_2=0.01. Watson is applied to every small data set with sample sizes 5.

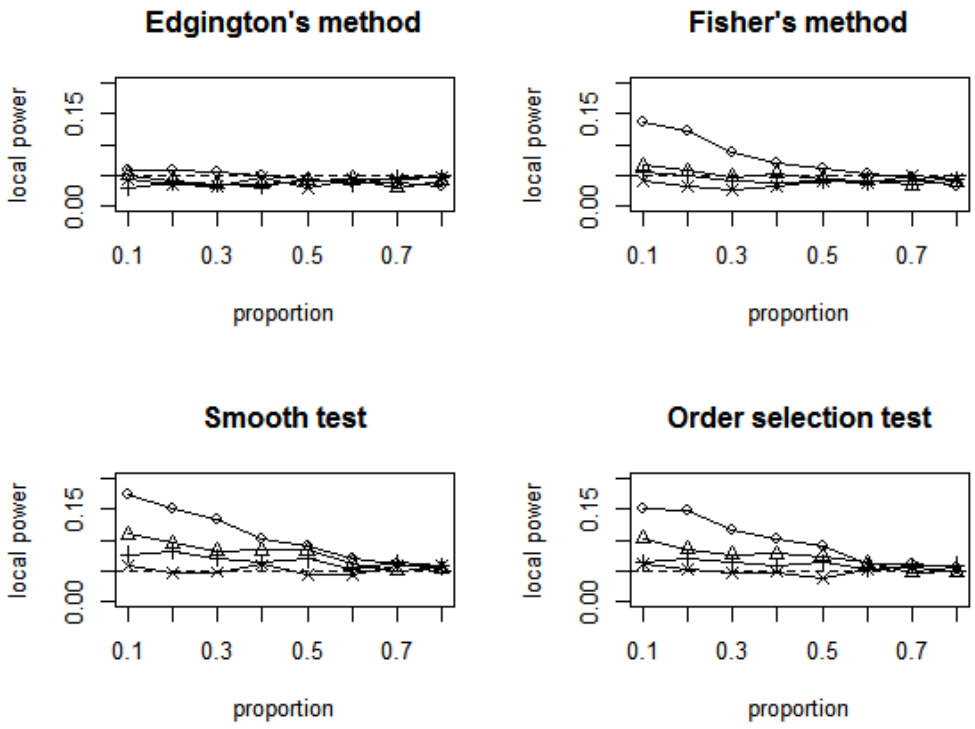

Figure 3.31: This figure shows the empirical power at the significance level 0.05 when testing whether data come from Laplace distributions, and the alternative is a mixture of Laplace and logistic distributions. The number of data sets considered are 100, 300, 500 and 1000, and the cross, plus, triangle and circle represent the number of data sets, respectively. For moment based tests, the two-sided test is used at the significance level α_2=0.01. AD is applied to every small data set with sample sizes 10.

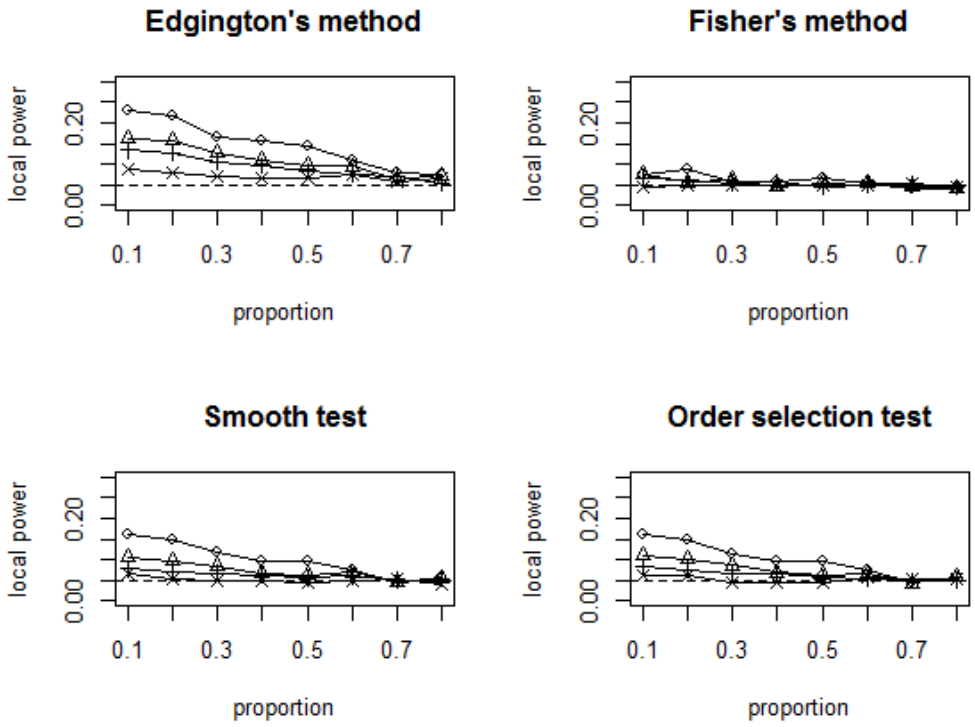

Figure 3.32: This figure shows the empirical power at the significance level 0.05 when testing whether data come from Laplace distributions, and the alternative is a mixture of Laplace and logistic distributions. The number of data sets considered are 100, 300, 500 and 1000, and the cross, plus, triangle and circle represent the number of data sets, respectively. Watson is applied to every small data set with sample sizes 10.

3.3 Testing whether data come from Weibull distributions

A Weibull distribution has been used in a variety of areas such as survival analysis, reliability analysis, and geophysics. For example, Carroll (2003) used Weibull distributions to analyze survival data from clinical trials and Heo et al. (2001) considered Weibull distributions for regional flood analysis. Other uses of Weibull distributions can be found in Chapter 3 of Pham (2006). One reason that a Weibull distribution is used in data analyses in a variety of areas is its flexibility due to the shape parameter. Hence, it may be natural to consider alternative distributions including a shape parameter. As alternative distributions, gamma distributions and log-normal distributions, are considered. These alternatives are selected because both distributions have the support of the positive real line and have a shape parameter.

There are two possible ways to test whether data come from Weibull distributions. One is to use the fact that the log-transformed Weibull distribution follows the Gumbel distribution, a location and scale family. The other is to use a test procedure which will be discussed in Chapter 5. The latter requires one to estimate the distribution of the shape parameter, and this step may cause an additional instability. Hence, the first way is preferable. To apply edf-based gof tests, estimates of location and scale parameters are necessary, and one of the most used estimators is the MLE. The MLE of location and scale parameters of Gumbel distributions are

$$\hat{\mu} = -\hat{\sigma} \log \left(\sum_{i=1}^{n} \exp\left(-\frac{x_i}{\hat{\sigma}}\right)/n \right)$$
$$\hat{\sigma} = \frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} x_i \exp\left(-x_i/\hat{\sigma}\right)}{\sum_{i=1}^{n} \exp\left(-x_i/\hat{\sigma}\right)}.$$

The MLE of the scale parameter must be found numerically, and this may not

be advisable under the current setting due to computational time and convergence issues. Other possible estimators are quantile-based estimators and moment estimators. To use quantile-based estimators, we need to choose appropriate quantiles. For example, Wang and Keats (1995) defined an improved quantile estimator of the shape parameter of Weibull distributions, and the estimator is based on an empirical quantile which minimized the bias of estimates in simulations. When we have data sets with small sample sizes, choosing an appropriate quantile like Wang and Keats (1995) is neither possible nor desirable, because we have few meaningful empirical quantiles due to the small sample size of each data set. Hence, it seems that the moment estimators are the most appropriate. Even if the moment estimators are not optimal in the mean squared error sense, they have a desirable property such as computational simplicity that is vital especially when we have a large number of data sets. The moment estimators for location and scale parameters of Gumbel distributions are $\hat{\mu} = \bar{x} - \gamma\hat{\sigma}$ where γ is Euler's constant and $\hat{\sigma} = \frac{\sqrt{6}}{\pi^2}s$ where s is the sample standard deviation. The location and scale-invariant property can be easily verified for these moment estimators.

Tables 3.25, 3.27 and 3.29 show the empirical size and power when all data sets come from the same distribution and the one-sided moment based tests, or smoothing based tests are applied. When the alternative is a log-normal distribution, both moment and smoothing based tests detect departures from the null well. Fisher's method based on AD has the highest power. When data sets are from gamma distributions, the power is relatively low. The power of smoothing based tests is just around the size of tests except when AD is used and we have 500 or 1000 data sets with 10 observations. Tables 3.26, 3.28 and 3.30 show the power when the two-sided moment based test is applied at the significance level α_2=0.01. Since there is no bias problem, the power decreases by the result of applying the two-sided

moment based tests. The effect of significance level α_2 is investigated in Figures 3.33 and 3.34 when we have 100 data sets with 5 observations. Under the log-normal alternative, Fisher's method based on P-values from AD attains the highest power at all considered significance levels α_2, and one interesting thing is that it has the least relative decrease in power. Under both alternatives, Edgington's method tends to have a bigger relative decrease in the power for each edf-based gof test.

Tables 3.31 to 3.36 show the local power, i.e., 90% of data sets come from the null distributions. When 10% of data sets come from log-normal distributions, moment based tests dominate smoothing based tests. Fisher's method based on P-values from AD attains the highest power. The power of both moment based tests and smoothing based tests is just around the size of tests when 10% of data sets come from gamma distributions. Figures 3.35 and 3.36 show the local power and the amount of relative decrease in the power at various significance levels α_2 when the two-sided moment based test is applied to 100 data sets with sample sizes 5. Since tests are not biased when 10% of data come from log-normal distributions, the power tends to decrease as the result of applying the two-sided moment based tests, and there is not much difference in the amount of relative decrease in the power for the different P-value combining methods. Under the gamma local alternatives, since the power is so low regardless of the type of edf-based gof test and the significance level α_2, it may be difficult to draw any meaningful conclusion. However, we notice that only when Fisher's method is applied to P-values from AD, the power is slightly greater than the size of the test. This may imply the preference of Fisher's method and AD.

Under both fixed and local alternatives, AD is the most powerful among the three edf-based gof tests. The performance of moment based tests and smoothing based tests depends on the alternative distribution. For example, when data are from log-normal distributions, moment based tests dominate smoothing based tests.

Especially, Fisher's method is better than Edgington's method when we have data sets with 5 observations. Both moment based and smoothing based tests are similar with respect to power when the data are from gamma distributions. Such results imply that applying Fisher's method to P-values from AD is desirable when we test whether data come from Weibull distributions.

Table 3.25: This table shows the size(%) and power(%) of the test. The null hypothesis is that data come from Weibull distributions and AD is applied to every small data set. For moment based tests, the one-sided test is used.

		Edgington's method			Fisher's method		
n	p	Weibull	Log-normal	Gamma	Weibull	Log-normal	Gamma
5	100	5.0	58.7	7.0	4.9	73.7	7.3
	300	5.2	94.8	8.8	4.9	98.3	9.6
	500	5.4	99.3	8.5	4.4	100.0	8.6
	1000	6.0	100.0	12.5	5.2	100.0	13.4
10	100	3.9	99.8	9.2	3.9	100.0	11.2
	300	4.0	100.0	12.1	4.2	100.0	15.9
	500	3.8	100.0	15.2	3.6	100.0	21.2
	1000	4.0	100.0	22.5	3.6	100.0	32.8
		Smooth Test			Order Selection Test		
n	p	Weibull	Log-normal	Gamma	Weibull	Log-normal	Gamma
5	100	5.6	46.1	5.6	5.4	44.9	4.8
	300	4.2	91.5	5.8	4.6	89.5	5.7
	500	5.1	99.1	5.8	4.5	98.8	5.1
	1000	5.5	100.0	7.4	6.1	100.0	7.0
10	100	5.2	99.1	7.1	5.1	99.4	6.0
	300	5.7	100.0	8.3	5.3	100.0	8.0
	500	5.2	100.0	9.3	5.3	100.0	8.2
	1000	5.6	100.0	14.2	5.7	100.0	13.0

Table 3.26: This table shows the size(%) and power(%) of the two-sided moment based test. The null hypothesis is that data come from Weibull distributions and AD is applied to every small data set. The significance level α_2=0.01 is used.

		Edgington's method			Fisher's method		
n	p	Weibull	Log-normal	Gamma	Weibull	Log-normal	Gamma
5	100	5.3	55.5	6.2	5.3	70.3	6.9
	300	4.8	93.5	7.5	4.8	98.0	8.2
	500	5.4	98.9	7.1	4.6	100.0	7.8
	1000	5.3	100.0	10.8	5.4	100.0	10.9
10	100	4.5	99.8	8.1	4.8	100.0	9.2
	300	4.8	100.0	10.8	4.5	100.0	13.9
	500	4.2	100.0	13.1	4.2	100.0	17.9
	1000	4.9	100.0	18.7	4.6	100.0	29.5

Table 3.27: This table shows the size(%) and power(%) of the test. The null hypothesis is that data come from Weibull distributions and CvM is applied to every small data set. For moment based tests, the one-sided test is used.

		Edgington's method			Fisher's method		
n	p	Weibull	Log-normal	Gamma	Weibull	Log-normal	Gamma
5	100	5.2	44.4	6.6	5.0	54.0	6.8
	300	5.2	83.2	7.6	5.1	91.4	8.0
	500	5.2	95.0	7.1	4.6	98.2	7.4
	1000	6.0	99.9	10.6	5.8	100.0	8.9
10	100	3.9	93.1	7.2	4.4	94.8	8.2
	300	4.2	100.0	8.9	4.4	100.0	10.0
	500	3.4	100.0	9.9	3.4	100.0	11.5
	1000	3.9	100.0	12.4	2.9	100.0	15.0
		Smooth Test			Order Selection Test		
n	p	Weibull	Log-normal	Gamma	Weibull	Log-normal	Gamma
5	100	5.3	26.4	5.1	5.8	29.1	4.6
	300	4.6	71.5	5.6	5.1	70.5	5.3
	500	4.6	90.5	5.3	4.6	88.6	4.8
	1000	5.6	99.8	6.8	5.6	99.4	6.3
10	100	5.1	81.3	6.0	5.3	85.2	5.5
	300	5.4	100.0	6.6	5.6	100.0	6.6
	500	5.7	100.0	6.6	5.4	100.0	6.3
	1000	5.8	100.0	7.3	5.7	100.0	7.1

Table 3.28: This table shows the size(%) and power(%) of the two-sided moment based test. The null hypothesis is that data come from Weibull distributions and CvM is applied to every small data set. The significance level α_2=0.01 is used.

		Edgington's method			Fisher's method		
n	p	Weibull	Log-normal	Gamma	Weibull	Log-normal	Gamma
5	100	5.3	39.9	5.8	5.0	49.9	6.3
	300	4.9	79.8	6.8	5.1	88.8	7.0
	500	5.4	94.0	6.7	4.8	97.6	6.2
	1000	5.3	99.8	9.4	5.6	100.0	8.6
10	100	4.7	91.3	6.6	4.5	93.8	7.8
	300	4.9	100.0	8.0	5.2	100.0	8.9
	500	3.8	100.0	9.0	4.0	100.0	9.8
	1000	4.5	100.0	10.3	4.1	100.0	11.8

Table 3.29: This table shows the size(%) and power(%) of the test. The null hypothesis is that data come from Weibull distributions and Watson is applied to every small data set. For moment based tests, the one-sided test is used.

		Edgington's method			Fisher's method		
n	p	Weibull	Log-normal	Gamma	Weibull	Log-normal	Gamma
5	100	5.1	33.4	6.2	5.2	36.8	6.6
	300	5.3	68.8	6.8	5.0	71.7	6.8
	500	5.2	85.2	6.4	4.3	88.7	6.2
	1000	5.8	98.1	9.2	5.9	99.0	7.4
10	100	4.2	73.4	6.2	4.4	70.2	6.8
	300	4.2	99.0	7.2	4.4	98.5	7.2
	500	3.4	99.9	7.3	3.2	100.0	7.7
	1000	3.6	100.0	8.1	3.1	100.0	8.6
		Smooth Test			Order Selection Test		
n	p	Weibull	Log-normal	Gamma	Weibull	Log-normal	Gamma
5	100	5.3	17.6	4.6	5.5	20.3	5.0
	300	4.4	51.1	5.3	5.1	52.7	4.9
	500	4.8	74.3	5.3	4.8	72.1	4.9
	1000	5.5	95.8	6.2	5.4	94.8	6.3
10	100	5.0	49.6	5.9	5.1	57.7	5.1
	300	5.5	96.5	5.8	5.7	96.4	6.0
	500	5.7	99.9	5.0	5.4	99.9	5.3
	1000	5.1	100.0	5.7	5.6	100.0	5.6

Table 3.30: This table shows the size(%) and power(%) of the two-sided moment based test. The null hypothesis is that data come from Weibull distributions and Watson is applied to every small data set. The significance level α_2=0.01 is used.

		Edgington's method			Fisher's method		
n	p	Weibull	Log-normal	Gamma	Weibull	Log-normal	Gamma
5	100	5.5	29.5	5.5	5.3	33.8	5.9
	300	5.1	64.8	6.1	5.1	67.8	6.0
	500	5.5	82.5	6.0	4.4	86.1	5.8
	1000	5.3	97.7	8.8	5.3	98.9	6.8
10	100	4.4	69.7	6.4	4.7	66.3	6.6
	300	4.9	98.9	6.8	5.0	98.2	6.7
	500	4.0	99.9	6.8	3.8	100.0	6.8
	1000	4.8	100.0	7.0	3.7	100.0	7.5

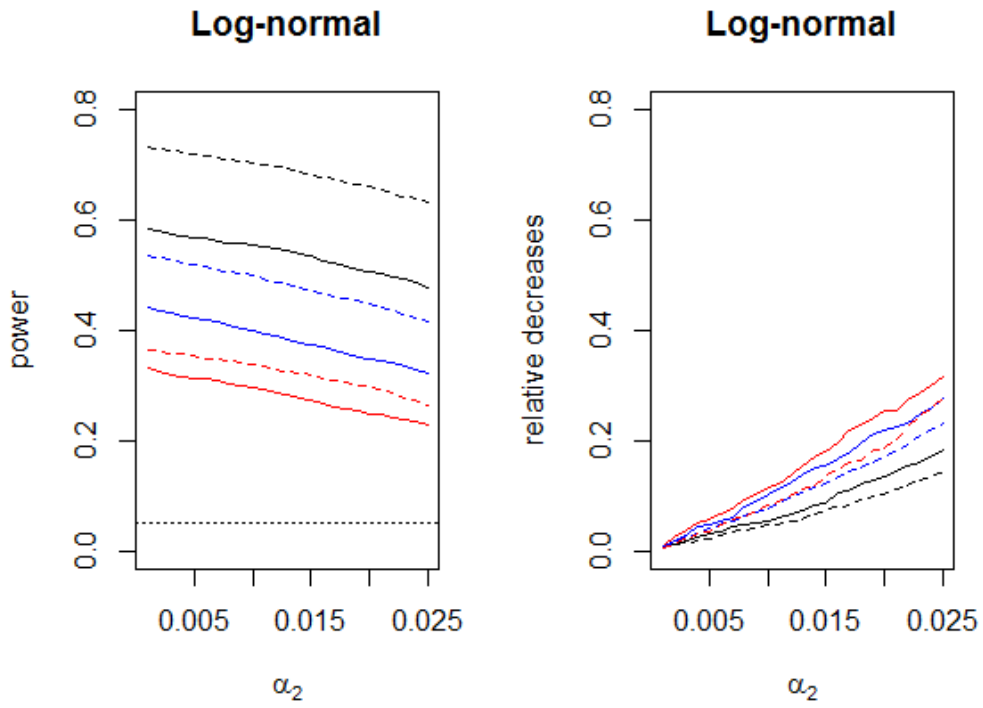

Figure 3.33: The left and right plots show the power of the two-sided moment based tests and the relative power decrease over various significance levels α_2, when there are 100 data sets with 5 observations. In both plots, black, blue and red lines represent AD, CvM, and Watson, respectively. Also, the solid and dashed lines represent Edgington's method and Fisher's method, respectively. The dotted line in the left plot denotes the significance level, 0.05.

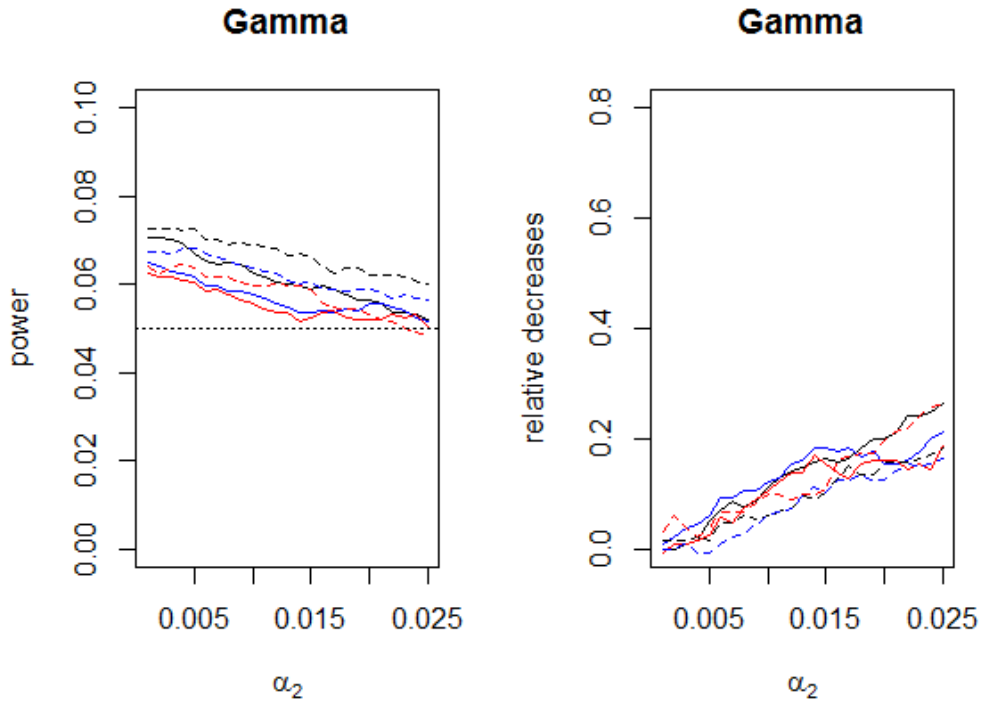

Figure 3.34: The left and right plots show the power of the two-sided moment based tests and the relative power decrease over various significance levels α_2, when there are 100 data sets with 5 observations. In both plots, black, blue and red lines represent AD, CvM, and Watson, respectively. Also, the solid and dashed lines represent Edgington's method and Fisher's method, respectively. The dotted line in the left plot denotes the significance level, 0.05.

Table 3.31: This table shows the local power(%) of the test when 90% of data sets are from the null distribution. The null hypothesis is that data come from Weibull distributions and AD is applied to every small data set. For moment based tests, the one-sided test is used.

		Edgington's method		Fisher's method		Smooth Test		Order Selection Test	
n	p	Log-normal	Gamma	Log-normal	Gamma	Log-normal	Gamma	Log-normal	Gamma
5	100	8.0	4.5	7.5	4.6	6.2	5.3	5.3	4.2
	300	9.3	6.0	10.1	6.3	6.3	5.4	6.2	5.1
	500	11.8	5.8	13.5	5.1	7.8	3.8	6.8	3.8
	1000	14.4	6.3	18.1	5.3	8.5	4.9	8.1	4.5
10	100	11.9	4.8	14.9	5.1	7.3	5.1	6.5	5.0
	300	17.9	4.4	27.8	4.6	11.3	4.4	9.5	4.5
	500	22.7	3.8	37.0	4.5	16.2	4.8	14.4	4.4
	1000	33.0	3.3	56.1	4.4	24.3	5.1	21.6	5.0

Table 3.32: This table shows the local power(%) of the two-sided moment based test when 90% of data sets are from the null distribution. The null hypothesis is that data come from Weibull distributions and AD is applied to every small data set. The significance level α_2=0.01 is used.

		Edgington's method		Fisher's method	
n	p	Log-normal	Gamma	Log-normal	Gamma
5	100	7.6	4.8	7.0	4.9
	300	8.5	6.0	8.3	6.1
	500	10.2	4.8	11.3	5.0
	1000	12.6	6.0	15.6	5.3
10	100	9.8	4.6	13.0	5.4
	300	15.4	4.4	24.0	4.9
	500	19.8	4.0	33.8	4.3
	1000	29.3	3.7	52.0	4.8

Table 3.33: This table shows the local power(%) of the test when 90% of data sets are from the null distribution. The null hypothesis is that data come from Weibull distributions and CvM is applied to every small data set. For moment based tests, the one-sided test is used.

		Edgington's method		Fisher's method		Smooth Test		Order Selection Test	
n	p	Log-normal	Gamma	Log-normal	Gamma	Log-normal	Gamma	Log-normal	Gamma
5	100	7.3	4.5	6.4	4.5	6.0	5.0	5.4	3.8
	300	8.4	6.2	8.5	5.8	5.6	5.8	5.5	5.1
	500	10.3	5.1	11.2	5.1	7.0	4.7	6.4	4.2
	1000	11.8	6.1	13.2	5.3	7.3	5.0	6.8	4.6
10	100	8.5	4.5	10.0	5.1	5.8	5.4	5.6	4.7
	300	11.8	3.9	15.0	4.2	6.8	4.8	6.1	4.5
	500	15.8	4.1	19.1	4.1	9.2	4.3	9.8	4.6
	1000	20.6	3.5	26.8	3.8	12.3	5.1	11.2	4.9

Table 3.34: This table shows the local power(%) of the two-sided moment based test when 90% of data sets are from the null distribution. The null hypothesis is that data come from Weibull distributions and CvM is applied to every small data set. The significance level α_2=0.01 is used.

		Edgington's method		Fisher's method	
n	p	Log-normal	Gamma	Log-normal	Gamma
5	100	6.4	4.3	6.0	5.1
	300	7.4	6.0	7.4	5.6
	500	8.6	4.3	9.2	4.6
	1000	9.9	5.6	11.2	5.1
10	100	7.5	4.5	8.4	5.3
	300	9.5	3.9	13.4	5.1
	500	13.6	4.2	16.7	4.2
	1000	18.1	3.8	23.1	4.8

Table 3.35: This table shows the local power(%) of the test when 90% of data sets are from the null distribution. The null hypothesis is that data come from Weibull distributions and Watson is applied to every small data set. For moment based tests, the one-sided test is used.

		Edgington's method		Fisher's method		Smooth Test		Order Selection Test	
n	p	Log-normal	Gamma	Log-normal	Gamma	Log-normal	Gamma	Log-normal	Gamma
5	100	7.0	4.6	6.2	4.6	5.4	5.1	5.8	3.8
	300	7.7	6.3	7.4	5.5	5.5	5.3	5.3	4.5
	500	8.8	5.5	9.3	5.0	6.6	4.7	5.9	4.7
	1000	10.1	6.2	10.9	5.3	7.0	4.6	6.5	4.7
10	100	7.4	4.5	7.3	5.3	5.1	5.0	5.4	4.7
	300	8.6	3.8	10.0	4.4	5.3	4.8	5.1	4.4
	500	11.6	4.0	11.5	4.3	7.0	4.8	7.6	5.0
	1000	13.1	3.3	13.3	3.5	8.2	5.7	7.6	5.3

Table 3.36: This table shows the local power(%) of the two-sided moment based test when 90% of data sets are from the null distribution. The null hypothesis is that data come from Weibull distributions and Watson is applied to every small data set. The significance level α_2=0.01 is used.

		Edgington's method		Fisher's method	
n	p	Log-normal	Gamma	Log-normal	Gamma
5	100	6.4	4.3	5.8	4.6
	300	6.5	5.8	6.8	5.3
	500	7.8	4.5	7.8	4.5
	1000	8.5	5.8	9.6	5.0
10	100	6.5	4.5	6.8	5.4
	300	7.3	3.6	8.6	5.0
	500	10.2	4.2	9.6	4.2
	1000	11.2	3.7	11.2	4.5

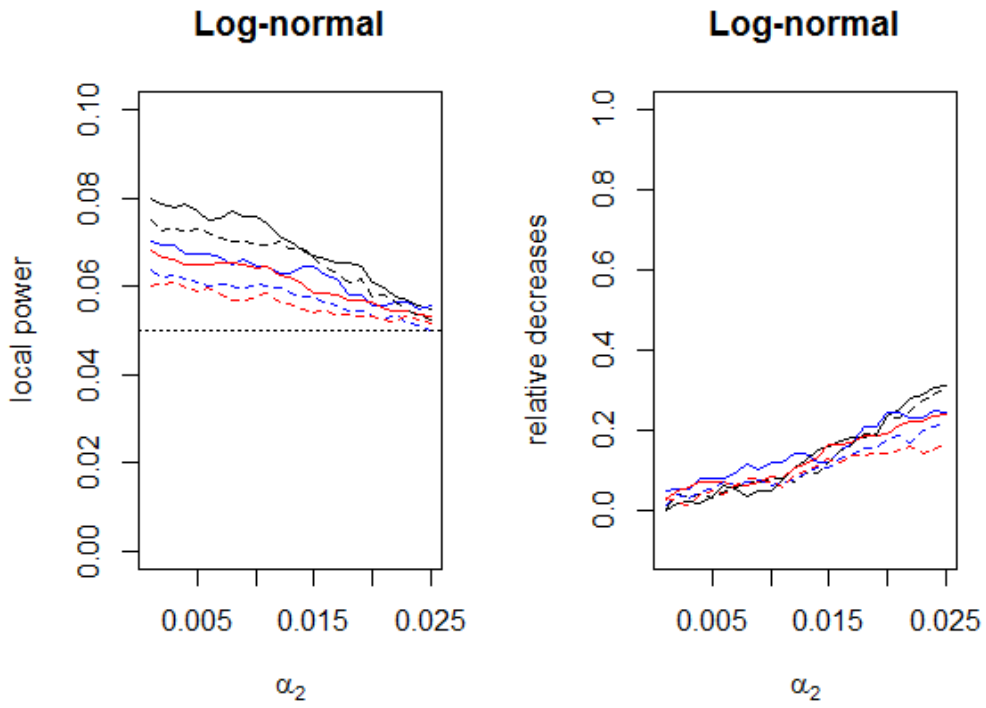

Figure 3.35: The left and right plots show the local power of the two-sided moment based tests and the relative power decrease over various significance levels α_2 when there are 100 data sets with 5 observations. In both plots, black, blue and red lines represent AD, CvM, and Watson, respectively. Also, the solid and dashed lines represent Edgington's method and Fisher's method, respectively. The dotted line in the left plot denotes the significance level, 0.05.

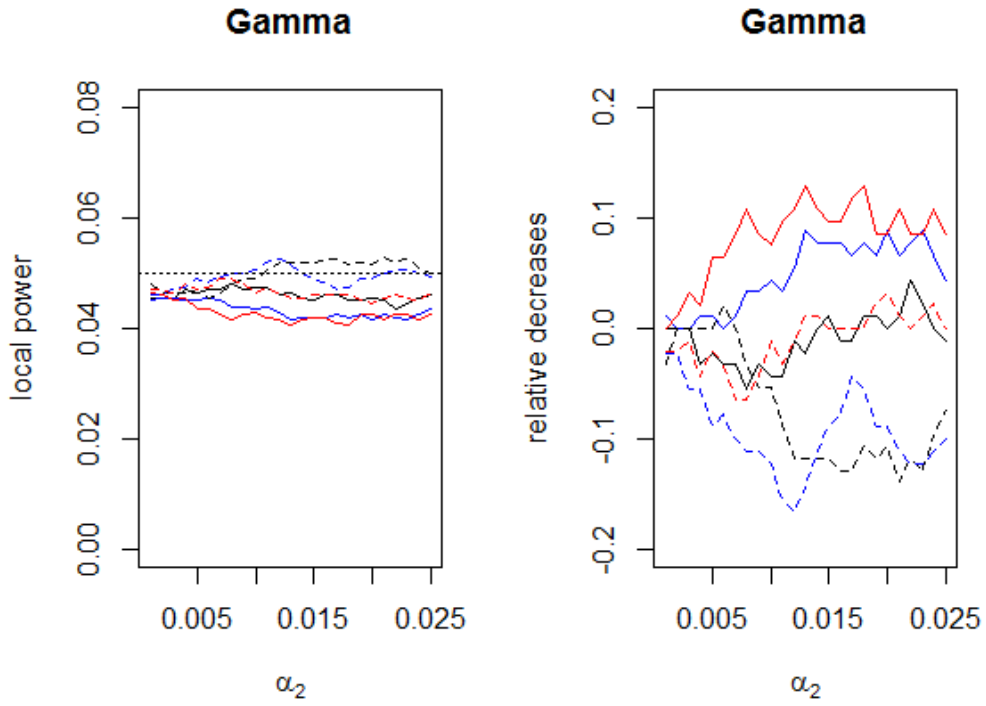

Figure 3.36: The left and right plots show the local power of the two-sided moment based tests and the relative power decrease over various significance levels α_2 when there are 100 data sets with 5 observations. In both plots, black, blue and red lines represent AD, CvM, and Watson, respectively. Also, the solid and dashed lines represent Edgington's method and Fisher's method, respectively. The dotted line in the left plot denotes the significance level, 0.05.

3.4 Summary of simulation results

We investigated the power of moment based and smoothing based tests via simulations, and found that the power depends on the considered null and alternative distributions, indicating that there does not exist a uniformly best method. When the null is normal, there does not exist much difference in power according to the type of edf-based gof test and Fisher's method tends to attain the highest power among the considered P-value combining methods. When we test whether data come from Laplace distributions and tests are not biased, moment based tests are better than smoothing based tests, and Edgington's method has slightly higher power. However, in this case, smoothing based tests still detect departures from the null well. When tests are biased, smoothing based tests are more powerful than moment based tests. Also, we noticed that two-sided moment based tests might not resolve the bias, especially when we have a relatively small number of data sets, such as 100 or 300. Of the three edf-based gof tests, Watson might be preferable to AD and CvM because AD seems to have the bias problem more frequently, and under the logistic local alternatives, CvM has power around the size for data sets of 5 observations, unlike AD and Watson. When the null is Weibull, under both fixed and local log-normal alternatives, moment based tests are better than smoothing based tests. Especially, Fisher's method based on P-values from AD is the most powerful. On the contrary, under the gamma local alternatives, both moment based and smoothing based tests have power close to the size of tests for all considered gof tests.

Since we do not have any information about a distribution from which data come, it is hard to choose one best method. However, according to the simulation results, if we consider the possible bias of tests, smoothing based tests based on P-values from Watson seems to be a safe choice.

4. REAL DATA EXAMPLE

In this chapter, we apply the suggested test procedure to microarray data collected by Robert Chapkin and coworkers at Texas A&M University. Previous analyses of the data set are found in Davidson et al. (2004), Hart and Cañette (2011) and Zhan and Hart (2012). Part of the data set, which contains 8038 logged gene expression levels from 5 rats, will be analyzed as in Hart and Cañette (2011) and Zhan and Hart (2012). Since Hart and Cañette (2011) found that there is strong evidence for scale differences between gene expression levels, we assume the following model for the data.

$$X_{ij} = \mu_i + \sigma_i \epsilon_{ij},\ i = 1, \cdots, 8038;\ j = 1, \cdots, 5.$$

We also assume that the errors are independent and identically distributed, and each ϵ_{ij} has mean 0 and variance 1. These assumptions follow Hart and Cañette (2011), and entail that the distributions of X_{ij} and $X_{lk} (i \neq l)$ differ only with respect to location and scale. We will consider two null distributions for ϵ_{ij}: normal and uniform. The uniform distribution is chosen because Hart and Cañette (2011) estimated error quantiles by the minimum distance method and found that they are remarkably close to uniform quantiles.

One important problem to be addressed before applying the test procedure to the data is possible correlations between logged gene expression levels in the same rat. Since the suggested test procedure is valid only when P-values from each small data set are independent, if there exist significant correlations between gene expression levels, applying the procedure to the data might result in poor power or incorrect size.

Fortunately, the independence assumption across genes was found to be reasonable by Zhan and Hart (2012) based on an analysis of autocorrelations.

Another problem is the number of bootstrap replications. Even if we checked that $100,000$ bootstrap replications are enough through simulation results in Chapter 3 when there are at most 1,000 data sets, there is a possibility that this number of bootstrap replications might not be enough when we have $8,038$ data sets. The number of bootstrap replications is especially important when we use Fisher's method because the sufficient condition for having the chi-squared null distribution for Fisher's method based on empirical P-values is $p = o(\sqrt{N})$ by Theorem 2.5.3. This implies that Edgington's method is better than Fisher's method when we have a large number of data sets. Also, it may be necessary to use a much larger number of bootstrap replications, such as 10^7, to generate the null distribution.

Since we assume that ϵ_{ij} are independent and identically distributed as a distribution with mean 0 and variance 1, we need to consider the uniform distribution on the interval $(-\sqrt{3}, \sqrt{3})$ when testing uniformity. Maximum likelihood is used to estimate the location and scale parameters, which are $\hat{\mu}_i = \dfrac{X_{i(1)} + X_{i(n)}}{2}$ and $\hat{\sigma}_i = \dfrac{X_{i(n)} - X_{i(1)}}{2\sqrt{3}}$, where $X_{i(j)}$ denotes the j-th order statistic within data set i. If AD is computed based on the MLE, AD always has the value ∞ regardless of the distribution from which observations come. This happens because the cumulative probabilities of the uniform distribution are always 0 and 1 for the smallest and largest observations, respectively. For this reason, AD is excluded when testing uniformity. One possible way to avoid excluding AD when testing uniformity is to use other estimators, perhaps moment estimators. Unfortunately, when testing uniformity, the moment estimator also has a problem in the sense that it does not guarantee that the inferred support includes all observations.

Table 4.1 shows test statistics and P-values of moment based tests and smoothing

Table 4.1: This table shows the test statistics and P-values of moment based tests and smoothing based tests regarding the number of bootstrap replications when testing whether data come from uniform distributions. The numbers in parentheses are the one-sided P-values.

Bootstrap Replications	CvM				Watson			
	Edgington	Fisher	Smooth	Order	Edgington	Fisher	Smooth	Order
10^5	6.82	15,165.78	65.18	42.52	7.93	15,165.90	74.21	64.78
	(1.00)	(1.00)	(6.7e-16)	(7.0e-11)	(1.00)	(1.00)	(0)	(8.9e-16)
10^6	6.80	15,159.44	65.42	46.28	8.00	15,154.85	72.67	67.24
	(1.00)	(1.00)	(5.6e-16)	(1.0e-11)	(1.00)	(1.00)	(0)	(2.2e-16)
10^7	6.84	15,153.26	65.92	45.30	7.98	15,169.88	75.08	65.96
	(1.00)	(1.00)	(4.4e-16)	(1.7e-11)	(1.00)	(1.00)	(0)	(4.4e-16)

based tests when testing uniformity. There does not exist much difference in test statistics depending on the number of bootstrap replications and the type of edf-based test statistics. When the one-sided moment based test is used, both Edgington's method and Fisher's method fail to reject the null hypothesis. However, if the two-sided moment based test is applied at the significance level α_2=0.001, uniformity is rejected. The results of the two-sided moment based tests agree with those of the smoothing based tests. The density estimate of the P-values in Figure 4.1 does not lie between the confidence bands, showing that the estimated density departs from uniformity. Also, both the moment based tests and the density estimate of the P-values imply that a relatively higher proportion of large P-values results in the rejection of the null hypothesis. This result conflicts with the usual belief that large P-values favor the null hypothesis and supports the idea of using the two-sided moment based tests instead of the one-sided ones.

Rejection of uniformity does not seem to accord with the estimated distribution of ϵ_{ij} from Hart and Cañette (2011). To explain this incompatibility, the following

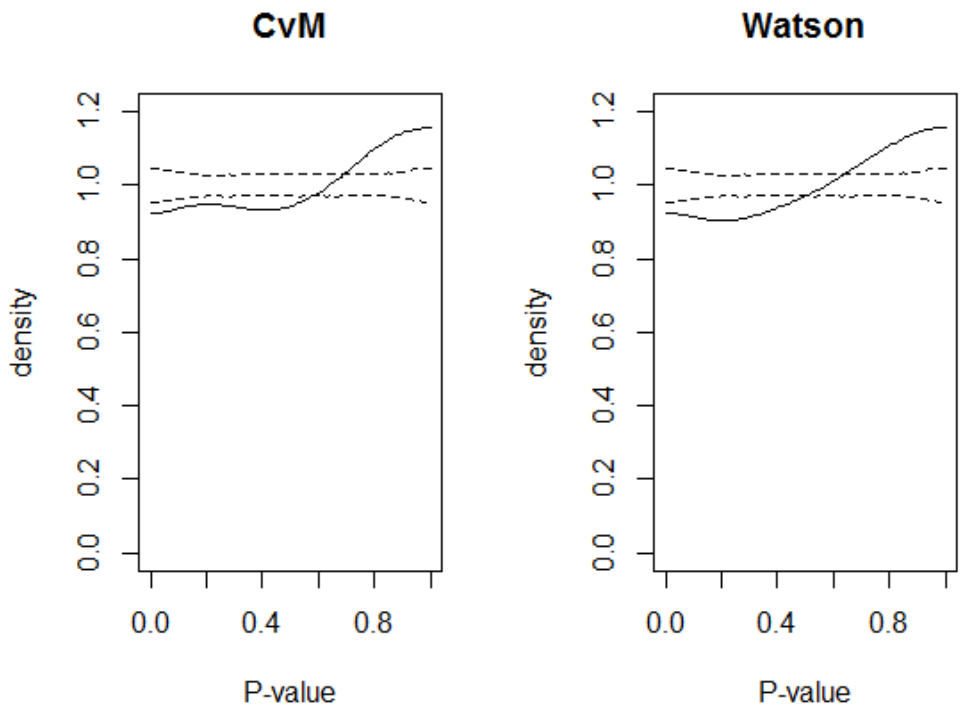

Figure 4.1: This figure shows the estimated density of P-values from each edf-based gof test when testing whether data come from uniform distributions. In these plots, P-values are obtained based on 10^7 bootstrap replications. The solid line represents the density estimate and the dashed lines represent 95% confidence bands for the density estimate when P-values are from the uniform distribution.

two alternative distributions are considered:

$$f_{10}(x) = \left(\frac{3 - 2\sqrt{3}}{6} x^2 + \frac{\sqrt{3} - 1}{2} \right) I_{(-\sqrt{3}, \sqrt{3})}(x)$$

$$f_{20,h}(x) = \sqrt{s_e^2 + h^2}\, \hat{f}_h(\sqrt{s_e^2 + h^2} x),$$

where $\hat{f}_h$ is a kernel density estimate using the Gaussian kernel and based on residuals $e_{ij} = \dfrac{X_{ij} - \bar{X}_i}{s_i}$, and h and s_e denote the bandwidth of the kernel estimate and the standard deviation of residuals, respectively. The density $f_{20,h}$ has mean 0 and

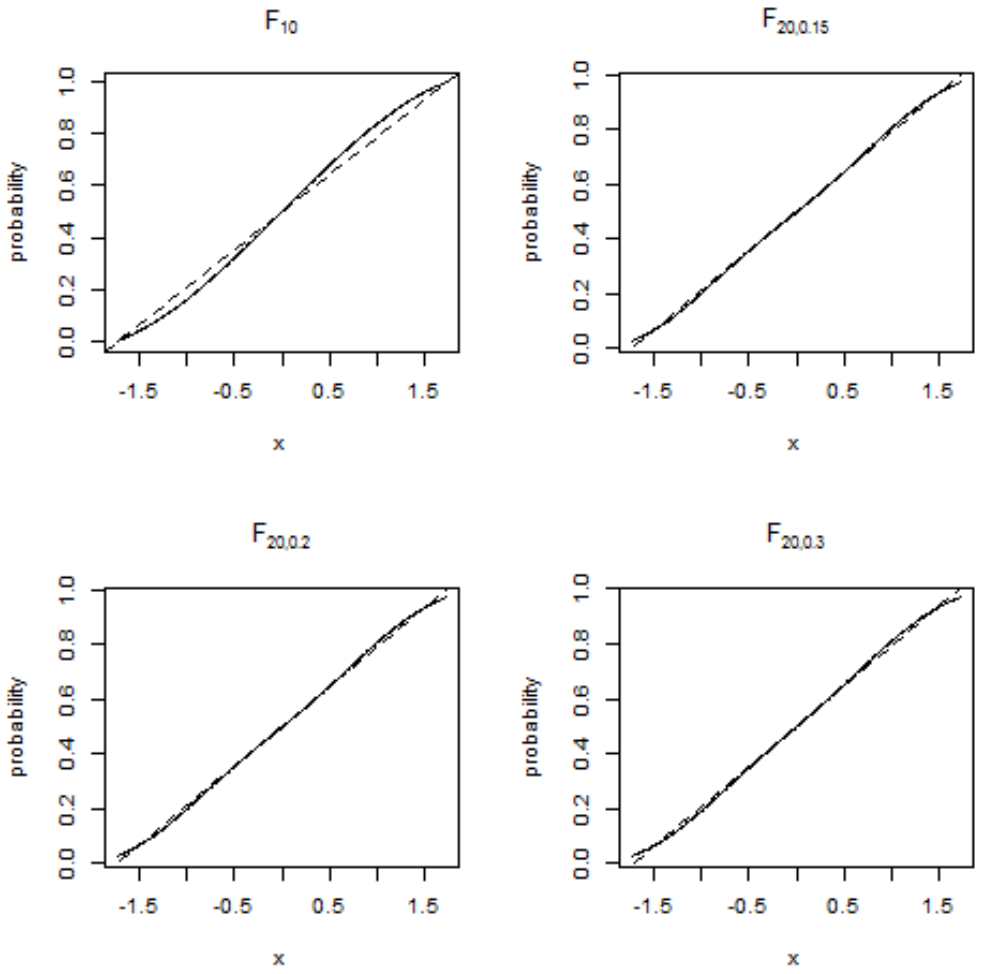

Figure 4.2: This figure shows the distribution functions of four alternatives. The solid and dashed lines represent the cumulative distribution functions of the alternative and null distribution, respectively.

variance 1 due to the facts that $E_{\hat{f}_h}(X) = \frac{1}{np}\sum_i\sum_j e_{ij}$ and $\mathrm{Var}_{\hat{f}_h}(X) = s_e^2 + h^2$. The distribution function of these alternatives are shown in Figure 4.2. Especially, when the alternative is $f_{20,h}$, we notice that there is little discrepancy between the cdf of the alternative distribution and that of the uniform distribution. This explains the apparent contradiction between our test result and the fact the quantile estimate

of Hart and Cañette (2011) appears very similar to that of a uniform distribution.

Tables 4.2 and 4.3 show the empirical size and power when testing whether data come from uniform distributions and CvM is applied to every small data set. To obtain the empirical P-values, 100,000 bootstrap replications were used. Since the results from Watson are similar to those from CvM, only results from CvM are shown here. Judging from the power of the one-sided moment based tests and the distribution of the P-value in Figure 4.3, we note that the test is biased under the considered alternatives, and the two-sided moment based tests and smoothing based tests detect departures from uniformity well. Since there exists little difference between the cdf of $f_{20,h}$ and the uniform cdf, the tests seem to well detect a very subtle difference between the null and the alternative, especially when we have a large number of data sets. Also, we notice that the size of Fisher's method is close to 0.10 when we consider 8,038 small data sets. The sufficient condition that Fisher's method based on empirical P-values has the chi-squared null distribution is $p = O(\sqrt{N})$ from Theorem 2.5.3, and this condition is not satisfied when there are 8,038 small data sets and just 100,000 bootstrap replications. Unfortunately, using the number of bootstrap replications satisfying the condition is too large and it is prohibitive to use too many bootstrap samples due to computing time. Hence, it may be desirable to use Edgington's method rather than Fisher's method when we have more than 1,000 data sets.

To check the hypothesis that the data set comes from a given alternative, our four tests are applied. The test statistics for Edgington's method, Fisher's method, the smooth test and the order selection test are -3.61, 16,769.07, 13.0, and 13.46 with P-values 0.0002, 6.8e-5, 0.0003, and 0.0002 when the null density is f_{10} and CvM is used. The results when Watson is used are similar to those when CvM is used. Clearly, both moment based tests and smoothing based tests show strong evidence

against the null density, f_{10}. To test whether the data set is from density $f_{20,h}$, we should not use the entire data set. If the whole data set is exploited, we use the same data set twice: it is used to obtain the kernel density estimate and to compute the test statistics. This indicates that the obtained results are not fair. Hence, the data set is randomly divided into half. Either of the two data sets can be used to obtain the kernel density estimate, and the test can be applied to the remaining one. To prevent the test results from depending on one random data split, we split the data set in two twenty times, and the results are shown in Table 4.4. When the bandwidth 0.15 is used, none of the randomly split data sets rejects the null. On the contrary, if either the bandwidth 0.2 or the bandwidth 0.3 is used, some reject the null. According to these results, it seems reasonable to conclude that $f_{20,0.15}$ is a good model for the distribution of the error density. We reiterate that this result is consistent with the estimated distribution of ϵ_{ij} by Hart and Cañette (2011) since Figure 4.2 shows that the distribution function of $f_{20,0.15}$ is close to that of the uniform distribution.

Tables 4.5 and 4.6 show the test statistics and P-values of moment based tests and smoothing based tests when testing normality. Both moment based tests using AD, CvM or Watson reject the null hypothesis regardless of whether one-sided tests or two-sided tests are used. Smoothing based tests also reject normality.

Results of testing uniformity and normality strongly suggest that the error density is short-tailed and if we are interested in testing whether the population means are 0, it would be better to use the linear signed rank test with scores designed for short-tailed densities rather than the t-test.

Table 4.2: This table shows the size(%) and power(%) of a nominal size 0.05 test. The null hypothesis is that data come from uniform distributions. CvM is applied to every small data set. For moment based tests, the one-sided test is used. Each value is obtained from 2,000 replications.

		Edgington's method					Fisher's method				
n	p	Uniform	f_{10}	$f_{20,0.15}$	$f_{20,0.2}$	$f_{20,0.3}$	Uniform	f_{10}	$f_{20,0.15}$	$f_{20,0.2}$	$f_{20,0.3}$
5	100	5.8	0.4	1.1	1.3	0.6	6.0	0.5	1.2	1.6	1.2
	300	5.7	0.0	0.4	0.1	0.0	5.4	0.1	0.6	0.3	0.4
	500	4.7	0.0	0.3	0.2	0.0	4.9	0.2	0.5	0.2	0.1
	1000	4.2	0.0	0.0	0.0	0.0	5.6	0.2	0.2	0.5	0.3
	8038	5.5	0.0	0.0	0.0	0.0	9.9	2.4	0.2	0.4	0.3

		Smooth Test					Order Selection Test				
n	p	Uniform	f_{10}	$f_{20,0.15}$	$f_{20,0.2}$	$f_{20,0.3}$	Uniform	f_{10}	$f_{20,0.15}$	$f_{20,0.2}$	$f_{20,0.3}$
5	100	5.4	12.9	7.0	8.6	8.6	5.0	16.9	9.0	10.4	13.4
	300	4.3	38.5	14.6	19.6	19.6	5.2	41.6	15.9	21.9	32.9
	500	5.1	57.7	25.1	31.4	31.4	5.1	59.8	26.2	32.5	50.6
	1000	5.2	89.0	42.9	55.4	55.4	4.2	89.0	42.1	54.7	79.0
	8038	5.3	100.0	100.0	100.0	100.0	5.3	100.0	100.0	100.0	100.0

Table 4.3: This table shows the size(%) and power(%) of the two-sided moment based tests. The null hypothesis is that data come from uniform distributions. CvM is applied to every small data set. The nominal size is 0.05 and the significance level α_2=0.01 is used. Each value is obtained from 2,000 replications.

		Edgington's method					Fisher's method				
n	p	Uniform	f_{10}	$f_{20,0.15}$	$f_{20,0.2}$	$f_{20,0.3}$	Uniform	f_{10}	$f_{20,0.15}$	$f_{20,0.2}$	$f_{20,0.3}$
5	100	6.1	10.9	5.4	7.0	8.2	6.4	9.9	4.9	6.3	7.0
	300	5.8	31.1	10.2	13.9	22.4	5.3	25.1	8.4	10.7	16.4
	500	4.3	48.9	17.2	24.7	38.7	4.8	37.5	12.8	15.7	27.6
	1000	4.8	83.3	32.7	43.0	70.0	5.7	71.0	23.9	29.6	52.6
	8038	5.3	100.0	32.7	43.0	70.0	9.3	100.0	100.0	100.0	100.0

Table 4.4: This table shows the percentage of rejections in 20 random splits of the data when we test whether the data come from $f_{20,h}$.

bandwidth	CvM				Watson			
	Edgington	Fisher	Smooth	Order	Edgington	Fisher	Smooth	Order
0.15	0	0	0	0	0	0	0	0
0.2	20.0	10.0	10.0	10.0	20.0	10.0	10.0	5.0
0.3	55.0	55.0	40.0	50.0	55.0	55.0	40.0	55.0

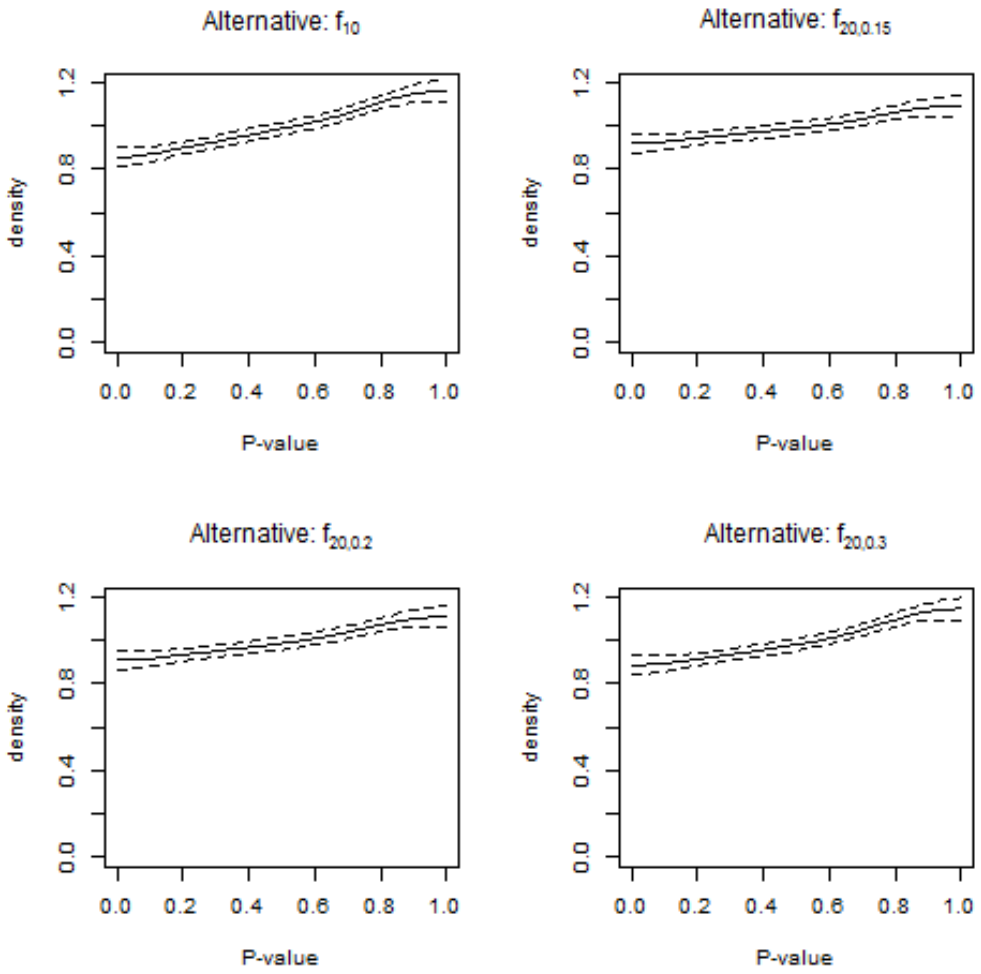

Figure 4.3: This figure shows the estimated density of P-values when testing uniformity and data come from alternative distributions. The solid line in each plot is the median of 1,000 kernel density estimates and the dashed lines are 0.025 and 0.975 percentiles of kernel density estimates.

Table 4.5: This table shows the test statistics and P-values of moment based tests regarding the number of bootstrap replications when testing whether the data set comes from normal distributions. The numbers in parentheses are the one-sided P-values.

Bootstrap Replications	AD		CvM		Watson	
	Edgington	Fisher	Edgington	Fisher	Edgington	Fisher
10^5	-7.44 (5.0e-14)	17,280 (2.7e-11)	-7.16 (1.9e-13)	17,160 (1.6e-9)	-7.61 (2.2e-13)	17,307 (1.0e-11)
10^6	-7.27 (1.8e-13)	17,232 (1.5e-10)	-6.98 (1.5e-12)	17,115 (6.8e-9)	-7.44 (5.0e-15)	17,264 (4.8e-11)
10^7	-7.25 (2.1e-13)	17,231 (1.5e-10)	-6.95 (1.8e-12)	17,112 (7.5e-9)	-7.41 (5.0e-14)	17,261 (5.4e-11)

Table 4.6: This table shows the test statistics and P-values of smoothing based tests regarding the number of bootstrap replications when testing whether the data set comes from normal distributions. The numbers in parentheses are the P-values.

Bootstrap Replications	AD		CvM		Watson	
	Smooth	Order	Smooth	Order	Smooth	Order
10^5	59.27 (1.4e-14)	49.69 (1.8e-12)	51.20 (8.3e-13)	46.60 (8.7e-12)	62.16 (3.2e-15)	52.10 (5.3e-13)
10^6	55.39 (9.9e-14)	48.86 (2.7e-12)	48.74 (2.9e-12)	45.89 (1.3e-11)	58.62 (1.9e-14)	51.19 (8.4e-14)
10^7	55.17 (1.1e-13)	51.67 (6.6e-13)	48.28 (3.7e-12)	48.72 (3.0e-12)	58.16 (2.4e-14)	54.20 (1.8e-13)

5. METHODOLOGY FOR NON-LOCATION AND SCALE FAMILY

When the null distribution is not in a location and scale family, the distributions of AD, CvM and Watson statistics depend on unknown parameters, indicating that the methodology in Chapter 2 cannot be applied. Several approaches have been proposed to deal with unknown parameters. One method is the half-sample method, which uses half of a data set to estimate the parameters, and computes gof test statistics based on the entire data set. This method asymptotically guarantees that the null distribution of test statistics when nuisance parameters are present is the same as that when nuisance parameters are not present. Stephens (1978) investigated the half-sample method when testing normality or exponentiality, and he found that there is considerable loss in power. Also, Braun (1980) suggested another method dealing with unknown parameters. The method randomly divides a data set into several groups and calculates gof test statistics for each group using estimates of parameters from all the observations. Each test statistic is compared to Bonferroni adjusted critical values and the null hypothesis is rejected when at least one test statistic is significant.

Clearly, both approaches might be used when we have a data set with a large enough sample size. When we have a large number of data sets with few replications, another approach to deal with unknown parameters is necessary. To handle the dependence on unknown parameters, we further assume that $\theta_1, \ldots, \theta_p$ are independent and identically distributed from a distribution G. We also assume that either there exists one unknown parameter or there are two unknown parameters, one of which is a location or scale parameter, to avoid the difficulty of dealing with multiple unknown parameters. In Section 5.1, the way to handle an unknown parameter will be

suggested. In Section 5.2, several issues which arise due to the existence of unknown parameters, such as the size of tests and the independence between P-values, will be discussed and simulation results will be shown when the null is a gamma distribution.

5.1 Estimating the distribution of unknown parameters and testing procedure

One way to handle the unknown parameter is to estimate its distribution and use the distribution to obtain unconditional P-values for every small data set. In the current setting, estimating the distribution of unknown parameters is equivalent to estimating a mixing distribution. Lindsay (1983) shows that for maximization purposes it is sufficient to consider a discrete measure with a finite set of positive probability, and the number of points of the support would not exceed the number of distinct data points. This implies that g, the density corresponding to G, can be estimated by a histogram-type estimator.

One issue related to this problem is identifiability of G. There is a literature exploring this issue (Teicher, 1961; Barndorff-Nielsen, 1965). Lindsay (1981) points out, however, that even if the mixing distribution G itself is not identifiable, there will be parameters of the mixture system which will be identifiable and estimable by the method of maximum likelihood. Hence, the mixing distribution, G, is either identifiable or not, but we can still estimate G using the maximum likelihood method. Also, since the purpose of estimating the density g is using it to obtain unconditional empirical P-values, identifiability of G does not matter.

Estimating the mixing distribution is a problem of maximizing the marginal likelihood of G. The only thing which is necessary to be found is the marginal likelihood. Suppose that $f(x;\alpha,\beta) = \frac{1}{\beta} f\left(\frac{x}{\beta};\alpha,1\right)$, i.e., distributions have a scale parameter and another parameter which is a shape parameter. Examples of such distributions are gamma and Weibull distributions. We also assume that $X_{i1},\ldots,X_{in}$ are inde-

pendent and identically distributed $f(\cdot|\alpha_i, \beta_i)$ given (α_i, β_i). It can be easily verified that the distribution of $U_i = \left(\frac{X_{i1}}{X_{in}}, \ldots, \frac{X_{i,n-1}}{X_{in}}\right), i = 1, \ldots, p$, depends only on the parameter α_i. Let h be the density function of U_i. Then the likelihood of a candidate $\tilde{G}$ for G is

$$L(\tilde{G}) = \prod_{i=1}^{p} \int h(U_i; \alpha) d\tilde{G}(\alpha).$$

We may model the density function g corresponding to distribution function G as

$$g(\alpha|\mathbf{p}) = \frac{k}{L} \sum_{j=1}^{k} p_j I_{\left(\frac{L(j-1)}{k}, \frac{Lj}{k}\right)}(\alpha),$$

where $L > 0$ is assumed to be such that $P(\alpha < L) \approx 1$, and $\mathbf{p} = (p_1, \ldots, p_k)$. Hence, the marginal log-likelihood of $\mathbf{p} = (p_1, \ldots, p_k)$ can be expressed as

$$l(\mathbf{p}) = \sum_{i=1}^{p} \log \left(\frac{k}{L} \sum_{j=1}^{k} p_j \int_{L(j-1)/k}^{Lj/k} h(u_i|\alpha) d\alpha \right). \tag{5.1}$$

The unconditional P-value of an observed test statistic t is $P(t) = P(T > t) = \int P(T > t|\alpha) dG(\alpha)$. Given an estimate $\hat{G}$ of G we may estimate the P-value by $\hat{P}(t) = \int P(T > t|\alpha) d\hat{G}(\alpha)$. Since the unconditional P-values $P(T_1), \ldots, P(T_p)$ are independent and identically distributed, and follow the uniform distribution under the null, it is reasonable to apply any one of the test procedures studied in Chapter 2. More specifically, we may proceed as follows:

1. For every small dataset, apply AD, CvM or Watson.

2. Estimate the distribution of the shape parameter by maximizing the marginal log-likelihood $l(\mathbf{p})$ in (5.1).

3. For some large number B, generate a sample $\alpha_1^*, \ldots, \alpha_B^*$ from $g(\alpha|\hat{\mathbf{p}})$.

4. Generate random samples of the same sample size as the data sets from $f(x; \alpha_i^*, 1)$, $i = 1, \ldots, B$, and compute AD, CvM or Watson for each of the B data sets.

5. Find the empirical unconditional P-values, $\frac{1}{B}\sum_{j=1}^{B} I(T_j^* > T_i)$ where T_j^* and T_i are test statistics from steps 4 and 1, respectively, and $i = 1, \ldots, p$.

6. Apply moment based tests or smoothing based tests to the empirical unconditional P-values from step 5.

When the null is from a location and scale family, to obtain empirical P-values we just need to generate bootstrap samples using the null distribution with location parameter 0 and scale parameter 1, because edf-based tests are invariant to location and scale parameters. When data are from a non-location and scale family, however, we need to generate bootstrap samples using the estimated density function g. This implies that there are two sources of error in this case: one is the error due to the finite number of bootstrap replications, and the other is the error due to estimating the density g. Hence, it is essential to find an appropriate number of bootstrap replications and number of bins. Since we use the whole data set to estimate the density g, one may also question whether unconditional P-values in step 5 are independent and identically distributed as the uniform distribution. These issues will be explored by an example of testing whether data come from gamma distributions in the next section.

5.2 Testing whether data come from gamma distributions

Suppose $X_1, \ldots, X_n$ are i.i.d with gamma distribution having shape parameter α and rate parameter β. Then the density of $\left(\frac{X_1}{X_n}, \ldots, \frac{X_{n-1}}{X_n}\right)$ is

$$f(y_1, \ldots, y_{n-1}|\alpha) = \frac{\Gamma(n\alpha)}{\Gamma(\alpha)^n}\left(\prod_{i=1}^{n-1} y_i\right)^{\alpha-1}\left(1 + \sum_{i=1}^{n-1} y_i\right)^{-n\alpha}, \text{ where } y_i = \frac{X_i}{X_n}. \quad (5.2)$$

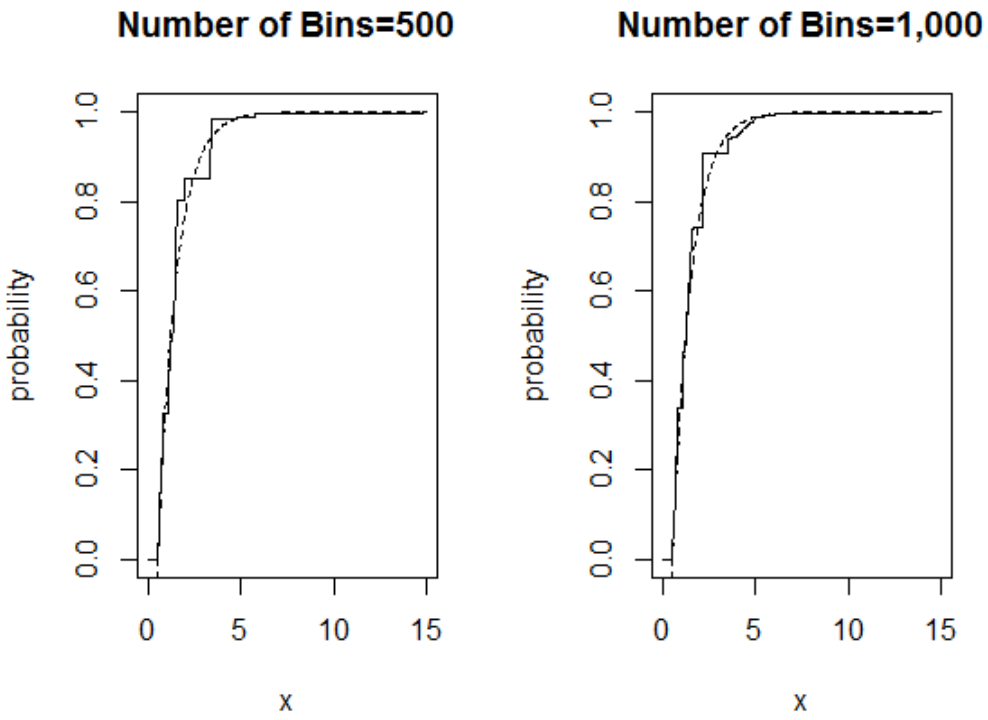

Figure 5.1: This figure shows the estimated distribution of the shape parameter when data come from gamma distributions. Shape parameters are generated from an exponential distribution with rate parameter 1, and then 1/2 is added. The left and right plots are the estimated distribution of the shape parameter for the number of bins, 500 and 1,000, respectively. In each plot, the solid and dotted lines represent the estimated distribution and the true distribution, respectively.

Using (5.2), we can now compute the likelihood of a candidate for g. To apply the test procedure in Section 5.1, we need to determine the number of bootstrap replications and bins. As a test case, we consider a situation when we have 1,000 data sets with 5 replications. The shape parameters of the gamma distributions are generated from an exponential distribution with a rate parameter 1, and then 1/2 is added to the obtained shape parameters. This distribution is selected because it

has a relatively low probability of having large values. Choosing such a distribution is important because D'Agostino and Stephens (1986, p.156) point out that, when testing whether data come from gamma distributions, critical values of AD or CvM do not change much for relatively large shape parameters. This implies that, when we have relatively large shape parameters, there is a possibility that tests may attain the right size even if the estimated density of the shape parameter is far from the true density. Since Lindsay (1981) shows that the number of bins need not exceed the number of data sets, three numbers of bins, 500, 750 and 1,000 will be considered.

Figure 5.1 shows the true and the estimated distributions of the shape parameter under 500 and 1,000 bins. For both numbers of bins, the estimated distribution is close to the true distribution. Tables 5.1 and 5.2 show the empirical size of tests at the significance level 0.05, for the different numbers of bins and replications. The size of tests tends to approach to the significance level as the number of bootstrap replication increases. Even if the number of replications is 15,000, however, the size of moment based tests is prone to be greater than 0.05 when the number of bins is 500 or 750. Only when the number of bootstrap replications is 10,000 or 15,000, and the number of bins is 1,000 do we obtain the right size for both moment based and smoothing based tests.

Another issue of interest is whether the obtained empirical P-values are independent of each other and are approximately uniformly distributed. Checking these might not be pragmatically necessary when the right size is obtained. However, since the right size does not guarantee independence and uniformity of empirical unconditional P-values, we will check these only for cases which have the right size. Specifically, we will check uniformity by the uniform Q-Q plot and tests of uniformity, and the independence will be checked by Hoeffding's independence test (Hoeffding, 1948).

Table 5.3 shows the results from testing uniformity at the significance level 0.05. The results do not seem to indicate departures from uniformity. Also, the uniform Q-Q plot in Figure 5.2 supports the uniformity of unconditional P-values from AD. We note that one P-value is randomly selected from each replication to draw the uniform Q-Q plot because the uniformity might be exaggerated by dependence if all P-values from the same replication were used. Even though the uniform Q-Q plots of P-values from CvM or Watson are not shown here, they also support uniformity. To perform the Hoeffding's independence test, 100 pairs of P-values are randomly selected from each replication. Specifically, 200 P-values are randomly selected from

Table 5.1: This table shows the size(%) of moment based tests according to the number of bins and replications for 1,000 data sets with sample sizes 5. Each value is obtained from 1,000 replications.

Bins	Replications	Edgington's method			Fisher's method		
		AD	CvM	Watson	AD	CvM	Watson
500	2,000	11.1	9.9	9.9	12.2	11.5	10.7
	4,000	9.0	7.2	7.3	8.8	8.2	8.1
	6,000	8.0	6.6	6.3	8.5	7.2	7.2
	8,000	7.6	6.6	6.4	8.5	7.0	7.0
	10,000	7.6	7.8	7.2	7.4	7.2	6.6
	15,000	7.2	6.0	6.2	6.0	6.4	6.8
750	2,000	10.1	10.3	10.6	11.4	10.9	10.8
	4,000	8.3	8.0	8.2	7.6	8.8	8.4
	6,000	6.4	6.6	6.4	6.7	6.9	6.7
	8,000	6.8	6.8	6.9	7.2	6.6	6.3
	10,000	6.1	5.4	4.9	5.6	5.8	5.3
	15,000	7.7	6.8	6.9	7.2	7.5	6.8
1,000	2,000	10.4	10.2	9.9	10.4	9.0	8.5
	4,000	7.3	7.4	7.4	8.7	8.3	8.3
	6,000	6.4	6.8	6.9	6.8	7.3	7.4
	8,000	5.9	5.5	5.4	5.8	6.4	6.2
	10,000	5.1	4.2	4.7	5.8	5.5	5.1
	15,000	5.1	4.4	4.5	6.7	5.6	5.2

each replication, and are divided in half. Then, the first P-value from each group is chosen to make a pair, and the remaining pairs are obtained in the same way. The results are shown in Table 5.4, and they indicate that the independence assumption is not violated.

Tables 5.5 to 5.10 show the empirical size and power of tests when we test whether data come from gamma distributions. In the simulation, two alternative distributions, log-normal distributions and Weibull distributions, are considered, and shape parameters are generated from an exponential distribution with a rate parameter 1, and then 1/2 is added. The distribution of the shape parameter is estimated by

Table 5.2: This table shows the size(%) of smoothing based tests according to the number of bins and replications for 1,000 data sets with sample sizes 5. Each value is obtained from 1,000 replications.

Bins	Replications	Smooth Test			Order Selection Test		
		AD	CvM	Watson	AD	CvM	Watson
500	2,000	13.4	13.9	13.6	13.7	14.6	14.7
	4,000	7.6	8.5	8.7	8.3	8.8	8.5
	6,000	8.5	8.9	8.8	8.1	8.3	8.0
	8,000	6.2	6.7	7.0	6.1	7.1	7.4
	10,000	5.8	5.8	6.0	6.6	6.0	6.0
	15,000	4.8	5.8	6.2	4.6	4.0	4.8
750	2,000	13.3	12.7	13.7	12.6	13.7	13.9
	4,000	8.9	9.4	9.7	8.5	9.2	8.9
	6,000	6.8	7.5	7.2	5.7	7.0	6.9
	8,000	6.6	6.1	6.1	6.0	5.8	5.5
	10,000	5.8	5.3	5.3	5.2	5.7	5.6
	15,000	5.4	6.4	6.0	5.7	6.5	6.4
1,000	2,000	16.0	12.5	12.4	11.6	11.4	13.9
	4,000	8.5	8.9	9.6	8.3	10.3	9.7
	6,000	7.1	6.7	6.6	7.0	6.6	6.4
	8,000	7.1	7.5	7.6	6.9	6.3	6.6
	10,000	4.6	5.4	5.8	5.0	5.5	5.4
	15,000	4.1	4.5	4.7	4.0	5.0	4.9

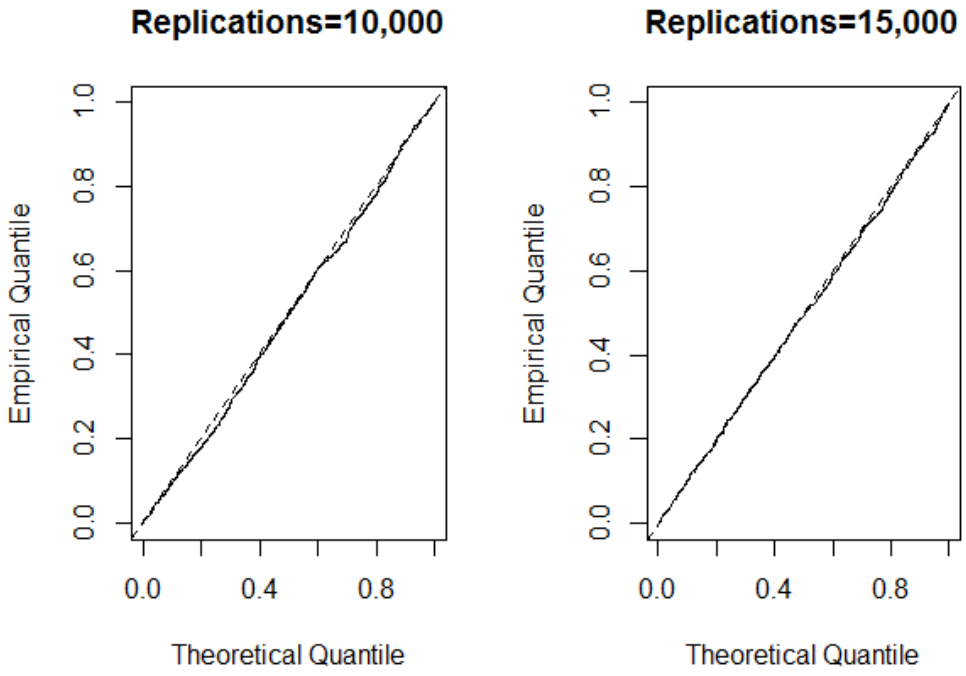

Figure 5.2: Both plots show the uniform Q-Q plot where the dashed line is a straight line with intercept 0 and slope 1.

using the number of bins equal to the number of data sets. Since both numbers of bootstrap replications, 10,000 and 15,000, seem to guarantee P-values that are independent and identically distributed as the uniform distribution, 10,000 bootstrap replications were used in the simulation to save computational time. In practice, if we have one data set, it may be better to use 15,000 bootstrap replications. We notice that, when moment based tests are applied, the size of tests is larger than the nominal level of 0.05. Even if we have 1,000 data sets with 10 observations, Fisher's method still fails to attain the right size. This indicates that more bootstrap replications are required for Fisher's method.

There exists a severe bias problem when the alternative is a log-normal distribution and moment based tests are applied to P-values from AD. One interesting thing

Table 5.3: This table shows the rejection percentage of testing uniformity of P-values from edf-based gof tests at the significance level $\alpha = 0.05$. Each value is obtained from 1,000 replications.

Bins	Replications	Smooth Test			Order Selection Test		
		AD	CvM	Watson	AD	CvM	Watson
1,000	10,000	4.6	5.4	5.8	5.0	5.5	5.4
	15,000	4.1	4.5	4.7	4.0	5.0	4.9

Table 5.4: This table shows the rejection percentage of Hoeffding's independence test based on 100 randomly selected pairs of P-values from AD, CvM or Watson at the significance level $\alpha = 0.05$. Each value is obtained from 1,000 replications.

Bins	Replications	AD	CvM	Watson
1,000	10,000	5.0	4.0	6.0
	15,000	4.0	3.0	5.0

is that moment based tests based on P-values from CvM or Watson do not have the bias problem. Also, it seems that the bias problem is resolved when the two-sided moment based tests are applied at the significance level α_2=0.01, except in the case of 100 data sets with sample sizes 5.

Figures 5.3 and 5.4 show the empirical power and relative decrease in power when the two-sided moment based tests are applied to 100 data sets with 5 observations at different significance levels α_2. When the alternative is a log-normal distribution, Fisher's method using P-values from CvM is the most powerful regardless of the significance levels α_2. In this case, the relative power decrease may be immaterial, but it still provides insights regarding a choice of the significance level α_2. For example, Figure 5.3 shows that at least 10% of power decreases when the significance level α_2 is greater than 0.009. One interesting thing is that, even if the power of Edgington's method based on AD is less than the nominal level of 0.05, it does not increase at

significance levels α_2 greater than 0.003. Especially, the bias problem of AD is not resolved even if an evenly divided significance level is used. When the alternative is a Weibull distribution, Fisher's method based on CvM has the highest power when the significance level α_2 is less than 0.019. On the contrary, Edgington's method based on AD has the best power when the significance level α_2 is greater than 0.021. Also, the power of Fisher's method tends to decrease relatively more than Edgington's method regardless of the type of gof tests, indicating that a cautious choice of the significance level α_2 may be more important for Fisher's method.

Tables 5.11 to 5.16 show the local power, i.e., 90% of data sets are from the null distribution, and the effect of the significance level α_2 is investigated in Figures 5.5 and 5.6. We notice that, when AD is applied, the power under the log-normal local alternatives is greater than that under the log-normal fixed alternatives. This result is surprising because it may not be expected that power increases when fewer data sets are from alternative distributions, and it happens due to the fact that AD is biased when data come from log-normal distributions, as shown in Table 5.5. Figure 5.5 shows that the power under the log-normal local alternatives is below the size when either Edgington's method or AD is used at some significance levels, such as 0.02. Also, we notice that the relative decrease in the power is the biggest when Fisher's method is applied to P-values from AD. These results may suggest CvM or Watson is preferable to AD.

Under the log-normal local alternatives, when CvM or Watson is used, Fisher's method attains the best power. However, under the Weibull local alternatives, there does not exist much difference in power according to the P-value combining methods. Such power results can be explained by the strength of evidence against the null. Figures 5.7 and 5.8 show the density estimate of the P-value when data sets are from log-normal distributions and Weibull distributions, respectively. The density of

the P-value, when Watson is applied, is not shown here, but it is similar to the density of the P-value when CvM is applied. From these figures, we notice that evidence against the null under the log-normal alternative is stronger than that under the Weibull alternative. Especially, when CvM is applied to data sets with the sample size 10, the evidence against the null is the strongest. This accords with higher power of Fisher's method than Edgington's method when CvM or Watson is applied to data sets that are from a mixture of gamma and log-normal distributions. According to the power results, when testing whether data come from gamma distributions, CvM or Watson is preferable to AD under both fixed and local alternatives. Also, moment based tests tend to have higher power than smoothing based tests.

Table 5.5: This table shows the size(%) and power(%) of a nominal size 0.05 test. The null hypothesis is that data come from gamma distributions and AD is used to compute the P-value. For moment based tests, the one-sided test is applied. Each value in the table is obtained from 1,000 replications.

		Edgington's method			Fisher's method		
n	p	Gamma	Log-normal	Weibull	Gamma	Log-normal	Weibull
5	100	8.4	2.8	11.4	9.4	0.1	9.2
	300	6.7	0.3	9.8	7.5	0.0	5.5
	500	7.1	0.0	11.2	7.0	0.0	4.5
	1000	5.1	0.0	12.5	5.8	0.0	4.5
10	100	9.1	44.7	19.0	11.7	28.0	14.8
	300	6.5	63.0	23.4	8.4	16.8	12.4
	500	6.2	78.6	30.5	6.8	17.8	14.8
	1000	5.3	93.2	47.0	6.4	18.3	18.2
		Smooth Test			Order Selection Test		
n	p	Gamma	Log-normal	Weibull	Gamma	Log-normal	Weibull
5	100	5.5	9.7	6.4	5.3	4.7	8.1
	300	4.7	45.5	6.0	4.9	28.9	7.5
	500	6.2	77.0	7.5	6.1	66.7	7.8
	1000	4.6	99.6	9.6	5.0	99.2	10.4
10	100	6.1	21.2	9.3	5.3	30.3	11.0
	300	5.5	43.2	12.3	5.0	50.3	14.9
	500	5.3	64.2	18.7	5.0	70.6	20.1
	1000	3.4	81.4	24.1	5.8	90.8	37.0

Table 5.6: This table shows the size(%) and power(%) of the two-sided moment based test. The null hypothesis is that data come from gamma distributions and AD is used to compute the P-value. The nominal size is 0.05 and the significance level α_2=0.01 is used. Each value in the table is obtained from 1,000 replications.

		Fisher's method			Edgington's method		
n	p	Gamma	Log-normal	Weibull	Gamma	Log-normal	Weibull
5	100	7.1	2.7	10.1	7.8	1.9	7.7
	300	5.5	5.6	8.6	6.1	34.7	5.4
	500	7.0	14.7	9.6	5.9	76.6	4.1
	1000	5.0	45.5	10.8	5.3	98.9	5.1
10	100	7.3	39.7	16.1	9.6	24.1	12.7
	300	6.0	57.9	21.0	7.3	12.7	10.7
	500	5.6	74.9	27.2	6.3	15.4	12.9
	1000	4.8	91.0	42.8	6.2	14.6	15.9

Table 5.7: This table shows the size(%) and power(%) of a nominal size 0.05 test. The null hypothesis is that data come from gamma distributions and CvM is used to compute the P-value. For moment based tests, the one-sided test is applied. Each value in the table is obtained from 1,000 replications.

		Edgington's method			Fisher's method		
n	p	Gamma	Log-normal	Weibull	Gamma	Log-normal	Weibull
5	100	5.7	33.9	11.5	6.4	61.7	13.2
	300	7.0	41.1	10.3	7.0	71.8	11.5
	500	5.9	51.3	12.6	6.1	83.7	12.8
	1000	4.2	72.3	16.8	5.5	95.5	18.3
10	100	8.2	96.0	17.2	8.5	99.7	23.1
	300	6.7	100.0	22.9	7.2	100.0	30.1
	500	5.7	100.0	28.4	7.0	100.0	37.3
	1000	6.2	100.0	42.5	5.9	100.0	54.9
		Smooth Test			Order Selection Test		
n	p	Gamma	Log-normal	Weibull	Gamma	Log-normal	Weibull
5	100	5.9	23.6	7.0	5.0	20.2	7.8
	300	5.4	30.0	8.7	5.3	27.1	9.1
	500	6.3	40.4	9.8	6.1	36.9	8.2
	1000	5.4	62.1	11.9	5.5	58.9	12.2
10	100	6.5	96.0	9.2	6.1	92.0	11.3
	300	5.4	100.0	13.6	5.4	100.0	14.4
	500	5.1	100.0	16.2	5.1	100.0	15.7
	1000	3.6	100.0	21.1	6.9	100.0	29.6

Table 5.8: This table shows the size(%) and power(%) of the two-sided moment based test. The null hypothesis is that data come from gamma distributions and CvM is used to compute the P-value. The nominal size is 0.05 and the significance level α_2=0.01 is used. Each value in the table is obtained from 1,000 replications.

		Edgington's method			Fisher's method		
n	p	Gamma	Log-normal	Weibull	Gamma	Log-normal	Weibull
5	100	5.6	29.2	10.0	6.2	57.7	11.7
	300	6.6	37.1	9.6	6.5	67.9	10.2
	500	6.0	46.5	10.7	5.6	81.3	11.0
	1000	5.0	67.2	15.1	6.5	95.5	18.6
10	100	7.3	94.8	14.9	7.4	99.7	20.0
	300	6.4	100.0	19.8	6.9	100.0	26.7
	500	5.4	100.0	25.6	6.5	100.0	34.1
	1000	6.4	100.0	38.5	7.0	100.0	54.9

Table 5.9: This table shows the size(%) and power(%) of a nominal size 0.05 test. The null hypothesis is that data come from gamma distributions and Watson is used to compute the P-value. For moment based tests, the one-sided test is applied. Each value in the table is obtained from 1,000 replications.

		Edgington's method			Fisher's method		
n	p	Gamma	Log-normal	Weibull	Gamma	Log-normal	Weibull
5	100	5.9	32.3	9.9	5.9	45.4	11.3
	300	7.0	43.8	9.2	6.6	60.5	9.4
	500	5.7	56.8	10.5	6.7	72.3	10.0
	1000	4.7	78.6	14.1	5.1	90.5	13.4
10	100	7.6	96.5	13.1	7.7	99.7	14.7
	300	6.2	100.0	16.4	7.0	100.0	16.8
	500	5.2	100.0	17.8	5.9	100.0	20.2
	1000	6.1	100.0	29.0	6.0	100.0	29.0
		Smooth Test			Order Selection Test		
n	p	Gamma	Log-normal	Weibull	Gamma	Log-normal	Weibull
5	100	6.0	19.1	6.9	4.5	19.7	6.7
	300	4.7	31.2	7.8	5.0	29.7	8.0
	500	5.5	44.5	9.3	5.6	43.2	7.8
	1000	5.8	68.6	10.5	5.4	67.5	10.4
10	100	6.2	94.6	7.7	5.9	91.7	8.0
	300	5.5	100.0	9.1	5.6	100.0	10.3
	500	5.3	100.0	10.5	4.9	100.0	9.9
	1000	3.3	100.0	11.1	7.1	100.0	18.50

Table 5.10: This table shows the size(%) and power(%) of the two-sided moment based test. The null hypothesis is that data come from gamma distributions and Watson is used to compute the P-value. The nominal size is 0.05 and the significance level α_2=0.01 is used. Each value in the table is obtained from 1,000 replications.

		Edgington's method			Fisher's method		
n	p	Gamma	Log-normal	Weibull	Gamma	Log-normal	Weibull
5	100	5.3	27.9	8.7	5.4	41.3	9.4
	300	6.8	39.7	9.2	6.4	56.5	8.3
	500	6.0	51.7	9.1	5.8	68.5	8.8
	1000	6.4	78.6	14.8	6.3	90.5	13.7
10	100	7.3	94.4	11.5	7.1	99.6	13.6
	300	6.5	100.0	13.6	6.5	100.0	14.8
	500	5.5	100.0	15.2	6.2	100.0	17.0
	1000	7.7	100.0	29.1	7.0	100.0	29.1

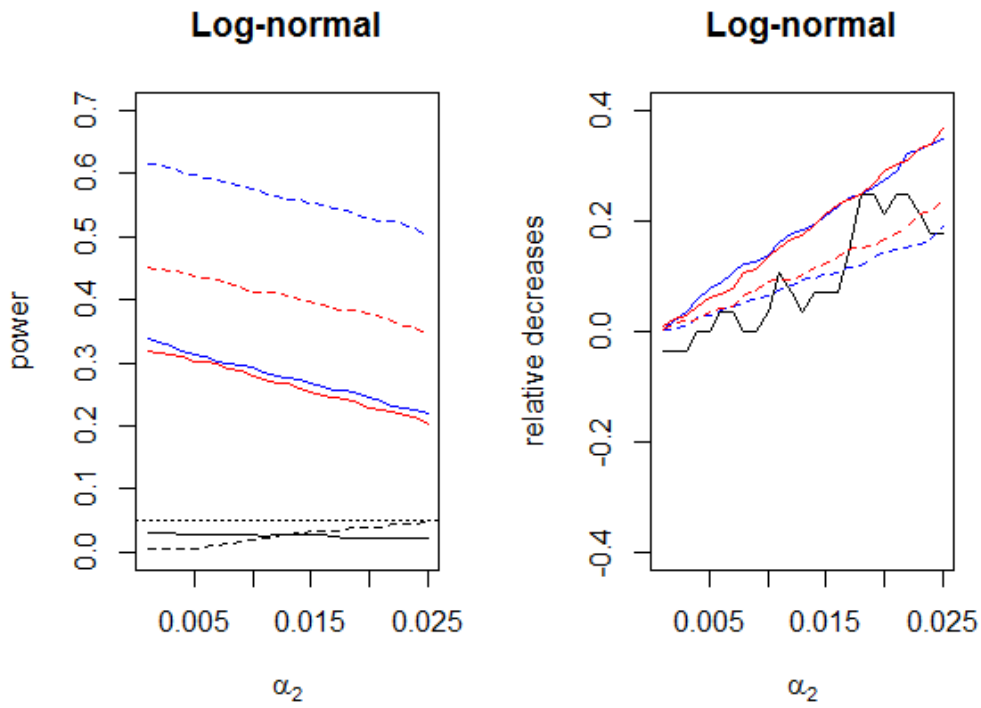

Figure 5.3: The left and right plots show the power of the two-sided moment based tests and the relative power decrease over various significance levels, α_2, respectively. In both plots, black, blue and red lines represent AD, CvM, and Watson, respectively. Also, the solid and dashed lines represent Edgington's method and Fisher's method, respectively. The dotted line in the left plot denotes the significance level 0.05.

Table 5.11: This table shows the local power(%) of a nominal size 0.05 test when 90% of data sets are from the null distributions. The null hypothesis is that data come from gamma distributions and AD is applied to every small data set. For moment based tests, the one-sided test is used.

		Edgington's method		Fisher's method		Smooth Test		Order Selection Test	
n	p	Log-Normal	Weibull	Log-Normal	Weibull	Log-Normal	Weibull	Log-Normal	Weibull
5	100	6.6	8.7	7.0	10.1	3.9	5.9	4.7	6.5
	300	5.3	6.5	3.4	5.8	5.6	5.5	5.3	3.9
	500	4.8	8.9	2.2	8.5	5.1	7.2	5.2	6.8
	1000	4.0	5.3	1.2	4.8	3.5	3.5	6.4	6.2
10	100	10.8	8.8	11.6	10.9	6.3	7.2	5.8	5.2
	300	9.1	8.4	7.1	7.7	6.5	5.7	5.7	6.0
	500	9.2	8.8	6.9	8.1	5.5	6.9	5.6	6.3
	1000	10.1	7.9	6.5	7.3	3.5	2.3	6.4	4.9

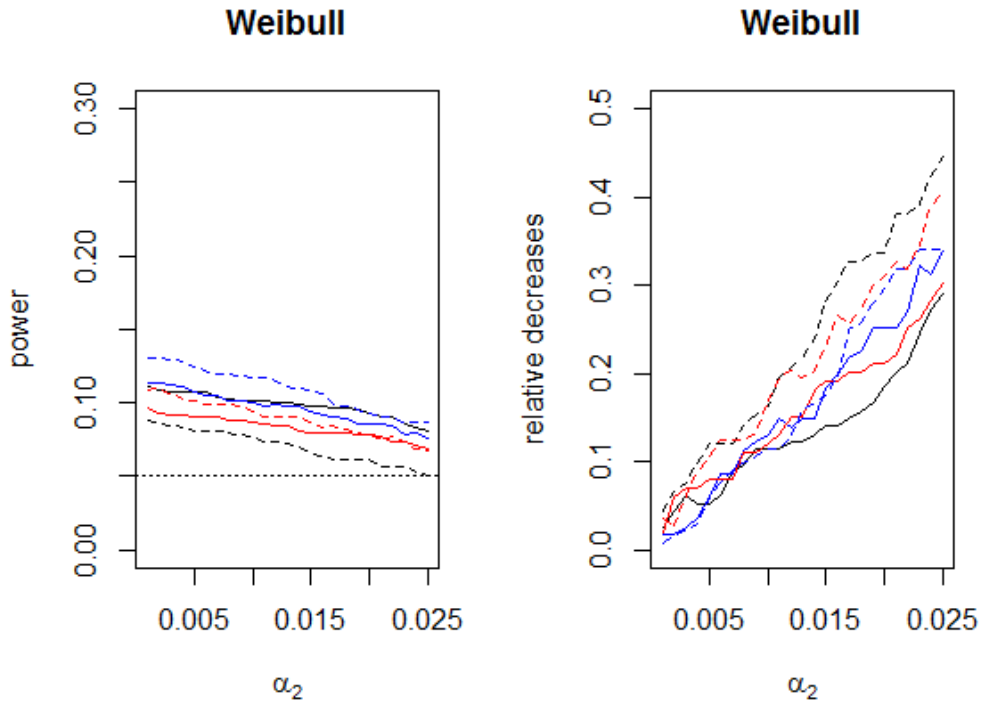

Figure 5.4: The left and right plots show the power of the two-sided moment based tests and the relative power decrease over various significance levels, α_2, respectively. In both plots, black, blue and red lines represent AD, CvM, and Watson, respectively. Also, the solid and dashed lines represent Edgington's method and Fisher's method, respectively. The dotted line in the left plot denotes the significance level 0.05.

Table 5.12: This table shows the local power(%) of two-sided moment based tests when 90% of data sets are from the null distributions. The null hypothesis is that data come from gamma distributions and AD is applied to every small data set. The nominal size is 0.05 and the significance level α_2=0.01 is used.

		Edgington's method		Fisher's method	
n	p	Log-Normal	Weibull	Log-Normal	Weibull
5	100	5.5	7.3	6.2	8.7
	300	5.2	5.9	3.9	5.3
	500	5.5	8.2	2.6	7.1
	1000	3.9	5.3	3.2	5.2
10	100	9.5	7.4	9.0	8.5
	300	8.0	7.8	5.9	7.0
	500	7.4	7.4	5.4	7.0
	1000	8.7	6.4	5.4	6.6

Table 5.13: This table shows the local power(%) of a nominal size 0.05 test when 90% of data sets are from the null distributions at the 5% significance level. The null hypothesis is that data come from gamma distributions and CvM is applied to every small data set. For moment based tests, the one-sided test is used.

		Edgington's method		Fisher's method		Smooth Test		Order Selection Test	
n	p	Log-Normal	Weibull	Log-Normal	Weibull	Log-Normal	Weibull	Log-Normal	Weibull
5	100	7.6	7.2	10.7	7.9	5.9	5.7	4.5	6.7
	300	8.6	6.5	11.4	6.9	5.8	5.4	6.3	4.9
	500	7.6	9.3	9.6	8.9	5.9	7.5	5.4	7.2
	1000	9.2	6.4	12.4	7.1	7.2	6.6	6.5	6.5
10	100	12.7	8.4	23.0	9.5	10.0	6.4	8.2	5.5
	300	14.3	7.4	31.6	7.6	13.2	6.2	8.9	5.0
	500	16.9	8.6	38.1	9.0	12.8	7.2	10.3	7.4
	1000	23.1	7.5	55.4	9.5	17.3	5.9	14.9	5.7

Table 5.14: This table shows the local power(%) of two-sided moment based tests when 90% of data sets are from the null distributions. The null hypothesis is that data come from gamma distributions and CvM is applied to every small data set. The nominal size is 0.05 and the significance level α_2=0.01 is used.

		Edgington's method		Fisher's method	
n	p	Log-Normal	Weibull	Log-Normal	Weibull
5	100	6.1	7.3	8.9	7.4
	300	7.5	5.9	10.2	5.8
	500	6.6	8.9	8.6	8.2
	1000	7.8	6.6	11.6	6.7
10	100	11.2	7.7	18.9	8.1
	300	12.7	6.8	27.8	6.5
	500	14.5	7.6	34.5	8.1
	1000	20.7	6.7	51.1	8.0

Table 5.15: This table shows the local power(%) of a nominal size 0.05 test when 90% of data sets are from the null distributions at the 5% significance level. The null hypothesis is that data come from gamma distributions and Watson is applied to every small data set. For moment based tests, the one-sided test is used.

		Edgington's method		Fisher's method		Smooth Test		Order Selection Test	
n	p	Log-Normal	Weibull	Log-Normal	Weibull	Log-Normal	Weibull	Log-Normal	Weibull
5	100	7.6	6.8	9.3	7.7	5.9	5.7	4.4	7.0
	300	8.7	5.9	9.8	6.6	6.4	5.3	6.6	4.8
	500	7.7	9.3	9.1	8.4	6.1	7.4	6.0	7.2
	1000	9.3	5.9	11.8	6.7	4.6	3.9	7.1	6.2
10	100	12.8	7.9	18.0	9.0	9.2	6.9	8.1	5.1
	300	15.1	6.8	28.0	6.6	12.3	6.1	9.5	5.0
	500	17.8	7.5	37.2	8.4	12.4	7.4	10.3	7.2
	1000	24.9	6.7	54.5	7.1	11.7	2.6	16.5	5.4

Table 5.16: This table shows the local power(%) of two-sided moment based tests when 90% of data sets are from the null distributions. The null hypothesis is that data come from gamma distributions and Watson is applied to every small data set. The nominal size is 0.05 and the significance level α_2=0.01 is used.

		Edgington's method		Fisher's method	
n	p	Log-Normal	Weibull	Log-Normal	Weibull
5	100	6.1	7.3	8.9	7.4
	300	7.5	5.9	10.2	5.8
	500	6.6	8.9	8.6	8.2
	1000	7.8	5.7	10.4	6.9
10	100	11.2	7.7	18.9	8.1
	300	12.7	6.8	27.8	6.5
	500	14.5	7.6	34.5	8.1
	1000	22.4	5.7	50.1	6.5

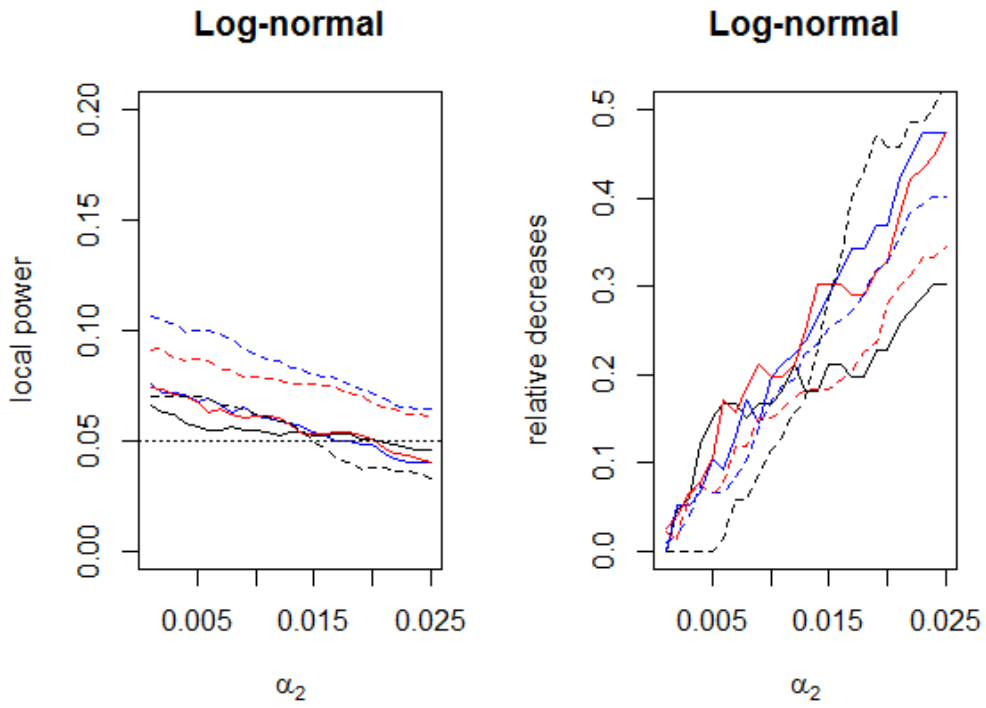

Figure 5.5: The left and the right plots show the power of the two-sided tests and the relative power decrease over various significance levels, α_2, respectively. In both plots, black, blue and red lines represent AD, CvM, and Watson, respectively. Also, the solid and dashed lines represent Edgington's method and Fisher's method, respectively. The dotted line in the left plot denotes the significance level, 0.05.

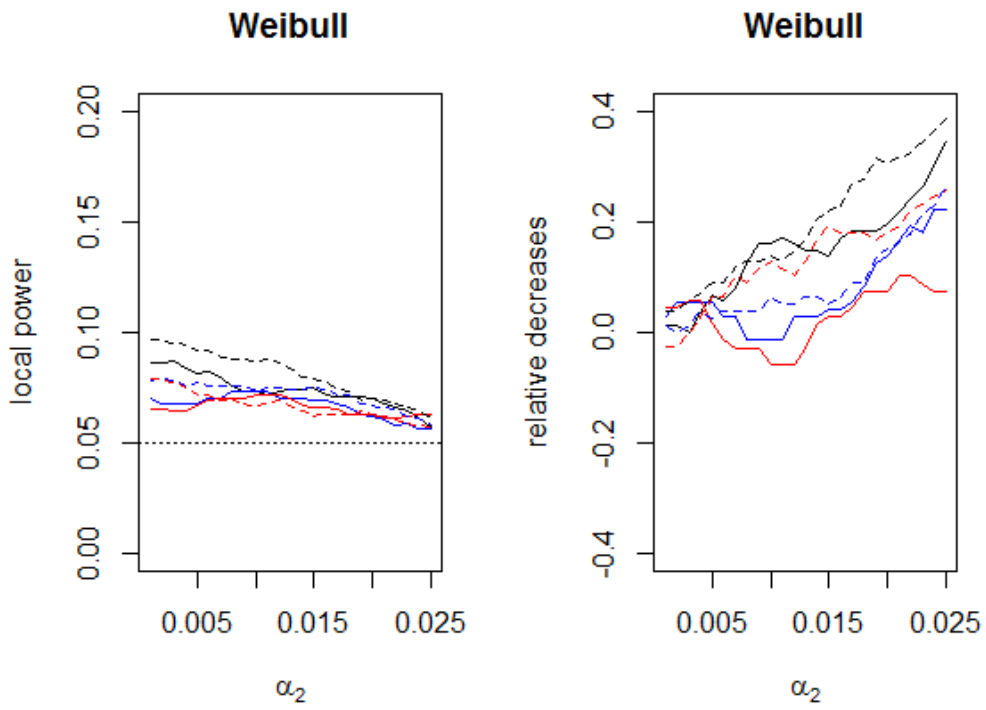

Figure 5.6: The left and the right plots show the power of the two-sided tests and the relative power decrease over various significance levels, α_2, respectively. In both plots, black, blue and red lines represent AD, CvM, and Watson, respectively. Also, the solid and dashed lines represent Edgington's method and Fisher's method, respectively. The dotted line in the left plot denotes the significance level, 0.05.

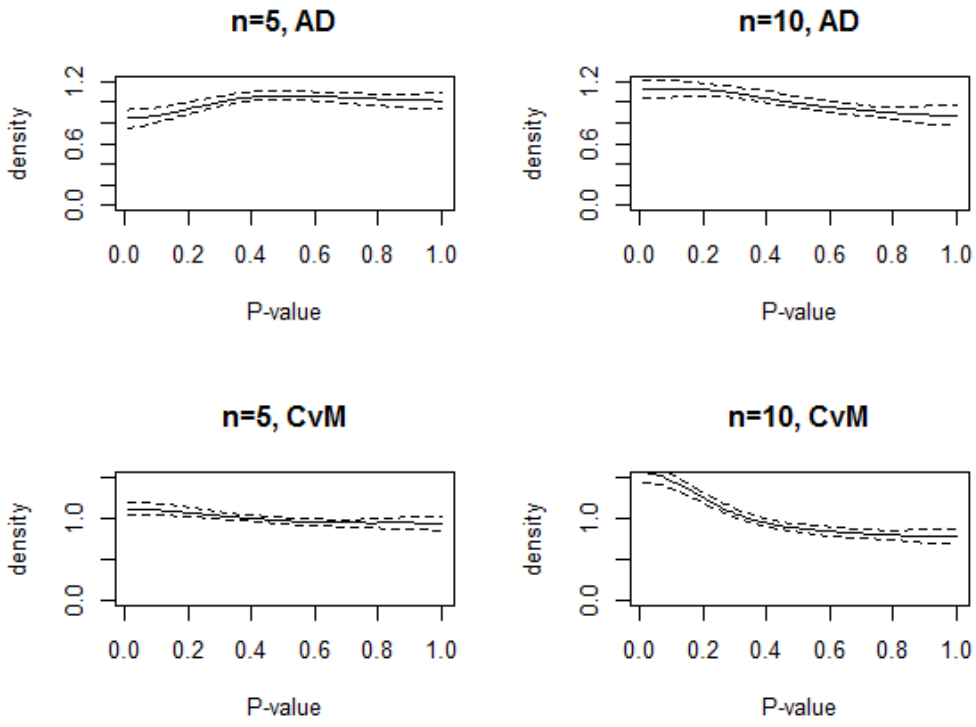

Figure 5.7: This figure shows the estimated density of P-values under the log-normal alternatives when testing whether data come from gamma distributions. In each plot, the solid line is the median of 100 kernel density estimates and the dashed lines are 0.025 and 0.975 percentiles of kernel density estimates.

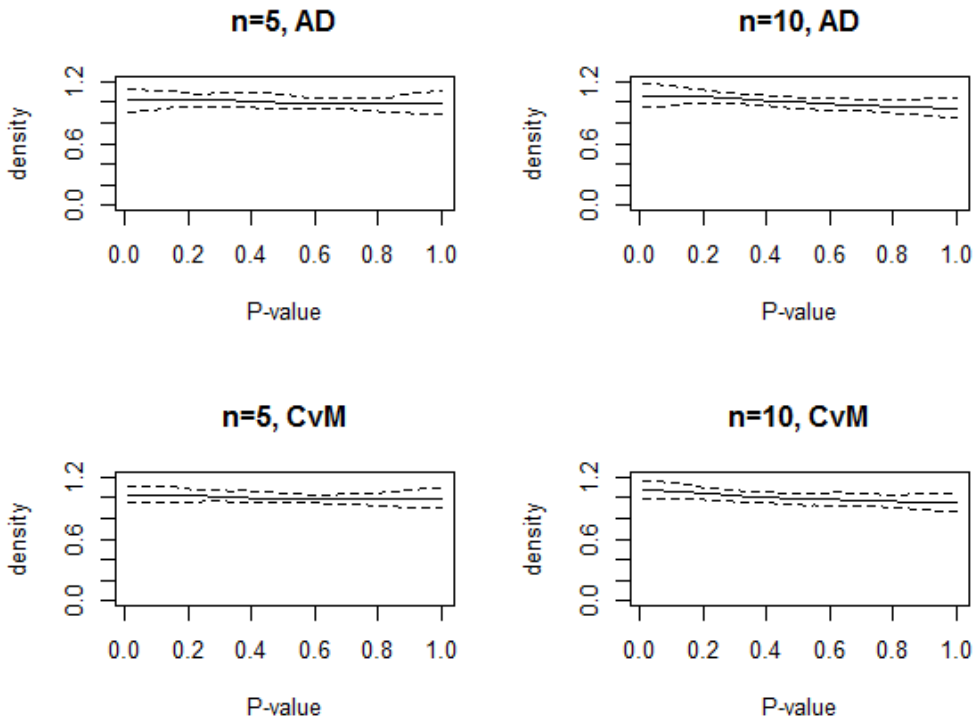

Figure 5.8: This figure shows the estimated density of P-values under the Weibull alternatives when testing whether data come from gamma distributions. In each plot, the solid line is the median of 100 kernel density estimates and the dashed lines are 0.025 and 0.975 percentiles of kernel density estimates.

6. SUMMARY AND FURTHER RESEARCH

6.1 Summary

In this dissertation, gof test procedures for a large number of small data sets are suggested and compared. The basic approach is to apply edf based gof tests to every small data set and use the fact that P-values follow the uniform distribution under the null. By exploiting uniformity, moment based tests or smoothing based tests can be applied to P-values to test whether all data sets come from a distribution in a specific parametric family. The two moment based tests, Edgington's method and Fisher's method, are compared regarding Pitman efficiency, and Edgington's method is shown to be slightly more efficient than Fisher's method. Also, for moment based tests, the two-sided test is suggested to handle possible bias due to small sample sizes. The effects of the two-sided tests are investigated under local alternatives at various significance levels α_2. These investigations indicate that it may be reasonable to use the significance level α_2 less than 0.015 at the significance level 0.05. Since the exact null distributions of edf based gof tests are unknown, we need to generate N bootstrap samples to obtain P-values. Conditions which guarantee that the asymptotic null distribution of moment based tests based on empirical P-values is the same as that based on exact P-values are found. For Edgington's method, the condition is $p = o(N)$, and for Fisher's method, the condition is $p = o(\sqrt{N})$.

When the null distribution is in a location and scale family, we can apply the suggested procedures easily because edf based gof tests are free of location and scale parameters. However, when the null distribution is not in a location and scale family, such as the gamma distribution, an additional step of estimating the distribution of an unknown parameter is required. The distribution of the unknown parameter can

be estimated by a histogram-type estimator. Since both the precision of the estimated density of the unknown parameter and the number of bootstrap replications may affect the fact that unconditional P-values are independent and asymptotically follow the uniform null distribution, effects of the number of bins and bootstrap replications are explored through the example of testing whether data come from gamma distributions. The example suggests that at least 10,000 bootstrap replications and the number of bins equal to the number of data sets are appropriate.

The power of moment based tests and that of smoothing based tests are investigated through simulations. Simulation results show that the two-sided moment based tests might not correct the bias problem, especially when we have a relatively small number of data sets, such as 100. Also, AD seems to suffer from a bias problem more frequently than CvM or Watson and Watson tends to have more stable power than CvM. These results suggest that using a smoothing based test based on Watson is desirable when we have a large number of data sets with few replications. Also, the suggested test procedures are applied to a real data set that has 8038 gene expressions from 5 mice. The real data analysis suggests that logged gene expression levels follow a short-tailed distribution, and if we need to perform a test about population means, it is better to use the linear signed rank test rather than the t-test.

6.2 Further Research

There are several possibilities for further research. We only found sufficient conditions for Fisher's method based on empirical P-values to have the chi-squared null distribution. The condition requires too many bootstrap replications and the simulation results in Chapter 3 indicate that we may need fewer bootstrap replications. Hence, finding necessary and sufficient conditions that guarantee the chi-squared null distribution for Fisher's method based on empirical P-values can be a part of further

research. Also, the test procedure suggested in this dissertation cannot be applied when we test whether data come from discrete distributions because the obtained P-values would be discrete and conservative. Thus, the test procedure is necessary to be modified to consider discrete null distributions, and this may be another area of further research.

REFERENCES

Allison, D. B., Gadbury, G. L., Heo, M., Fernández, J. R., Lee, C.-K., Prolla, T. A., and Weindruch, R. (2002), "A mixture model approach for the analysis of microarray gene expression data," *Computational Statistics & Data Analysis*, 39, 1–20.

Arshad, M., Rasool, M., and Ahmed, M. (2002), "Power study for empirical distribution function tests for generalized Pareto distribution," *Pakistan Journal of Applied Science*, 2, 1119–1122.

Barndorff-Nielsen, O. (1965), "Identifiability of mixtures of exponential families," *Journal of Mathematical Analysis and Applications*, 12, 115–121.

Berk, R. and Cohen, A. (1979), "Asymptotically Optimal Methods of Combining Tests," *Journal of the American Statistical Association*, 74, 812–814.

Birnbaum, A. (1954), "Combining independent tests of significance," *Journal of the American Statistical Association*, 49, 559–574.

Braun, H. (1980), "A simple method for testing goodness of fit in the presence of nuisance parameters," *Journal of the Royal Statistical Society. Series B*, 53–63.

Cao, R. and Lugosi, G. (2005), "Goodness-of-fit Tests Based on the Kernel Density Estimator," *Scandinavian Journal of Statistics*, 32, 599–616.

Carroll, K. J. (2003), "On the use and utility of the Weibull model in the analysis of survival data," *Controlled clinical trials*, 24, 682–701.

Cohen, A., Marden, J., and Singh, K. (1982), "Second order asymptotic and non-asymptotic optimality properties of combined tests," *Journal of Statistical Planning and Inference*, 6, 253–276.

Cox, D. and Solomon, P. (1986), "Analysis of Variability with large numbers of small samples," *Biometrika*, 73, 543–554.

Dadi, M. and Marks, R. (1987), "Detector relative efficiencies in the presence of Laplace noise," *IEEE transactions on aerospace and electronic systems*, 568–582.

D'Agostino, R. B. and Stephens, M. (1986), *Goodness-of-fit-techniques*, vol. 68, New York: Marcel Dekker, INC.

Davidson, L. A., Nguyen, D. V., Hokanson, R. M., Callaway, E. S., Isett, R. B., Turner, N. D., Dougherty, E. R., Wang, N., Lupton, J. R., Carroll, R. J., et al. (2004), "Chemopreventive n-3 polyunsaturated fatty acids reprogram genetic signatures during colon cancer initiation and progression in the rat," *Cancer Research*, 64, 6797–6804.

Easterling, R. G. (1978), "Exponential responses with double exponential measurement error-A model for steam generator inspection," in *Proceedings of the DOE Statistical Symposium, US Department of Energy*, pp. 90–110.

Edgington, E. (1972), "An additive method for combining probability values from independent experiments," *Journal of Psychology*, 80, 351–363.

Fan, Y. (1994), "Testing the Goodness of Fit of a Parametric Density Function by Kernel Method," *Econometric Theory*, 10, 316–356.

— (1998), "Goodness-of-Fit Tests Based on Kernel Density Estimators with Fixed Smoothing Parameters," *Econometric Theory*, 14, 604–621.

Filliben, J. J. (1975), "The probability plot correlation coefficient test for normality," *Technometrics*, 17, 111–117.

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Edinburgh and London: Oliver & Boyd.

Frain, J. C. (2007), "Small sample power of tests of normality when the alternative is an α-stable distribution," *Trinity Economics Papers TEP-0207, Trinity College Dublin, Department of Economics*.

Gel, Y. R. (2010), "Test of fit for a Laplace distribution against heavier tailed alternatives," *Computational Statistics & Data Analysis*, 54, 958–965.

Green, J. and Hegazy, Y. (1976), "Powerful modified-EDF goodness-of-fit tests," *Journal of the American Statistical Association*, 71, 204–209.

Greenwood, M. (1946), "The statistical study of infectious diseases," *Journal of the Royal Statistical Society*, 109, 85–110.

Gürtler, N. and Henze, N. (2000), "Goodness-of-fit tests for the Cauchy distribution based on the empirical characteristic function," *Annals of the Institute of Statistical Mathematics*, 52, 267–286.

Hart, J. (1997), *Nonparametric Smoothing and Lack-of-Fit Tests*, New York: Springer, 1st ed.

Hart, J. D. and Cañette, I. (2011), "Nonparametric estimation of distributions in random effects models," *Journal of Computational and Graphical Statistics*, 20, 461–478.

Heo, J.-H., Boes, D., and Salas, J. (2001), "Regional flood frequency analysis based on a Weibull model: Part 1. Estimation and asymptotic variances," *Journal of hydrology*, 242, 157–170.

Hoeffding, W. (1948), "A non-parametric test of independence," *The annals of mathematical statistics*, 546–557.

Hsu, D. (1979), "Long-tailed distributions for position errors in navigation," *Applied Statistics*, 62–72.

Inglot, T. and Ledwina, T. (2006), "Towards data driven selection of a penalty function for data driven Neyman tests," *Linear Algebra and its Applications*, 417, 124–133.

Jarque, C. M. and Bera, A. K. (1980), "Efficient tests for normality, homoscedasticity and serial independence of regression residuals," *Economics letters*, 6, 255–259.

Kallenberg, W. C. M. and Ledwina, T. (1997), "Data-Driven Smooth Tests when the Hypothesis Is Composite," *Journal of the American Statistical Association*, 92, 1094–1104.

Kim, J. T. (2000), "An order selection criterion for testing goodness of fit," *Journal of the American Statistical Association*, 95, 829–835.

Kotz, S., Kozubowski, T. J., and Podgórski, K. (2001), "Asymmetric multivariate laplace distribution," in *The Laplace Distribution and Generalizations*, Springer, pp. 239–272.

Kyriakoussis, A., Li, G., and Papadopoulos, A. (1998), "On characterization and goodness-of-fit test of some discrete distribution families," *Journal of Statistical Planning and Inference*, 74, 215–228.

Lancaster, H. (1961), "The combination of probabilities: an application of orthonormal functions," *Australian Journal of Statistics*, 3, 20–33.

Ledwina, T. (1994), "Data-Driven Version of Neyman's Smooth Test of Fit," *Journal of the American Statistical Association*, 89, 1000–1005.

Liang, J., Tang, M., and Chan, P. S. (2009), "A generalized Shpiro-Wilk W statistic for testing high-dimensional normality," *Computational Statistics and Data Analysis*, 53, 3883–3891.

Lindsay, B. G. (1981), "Properties of the maximum likelihood estimator of a mixing

distribution," in *Statistical Distributions in Scientific Work*, Springer, pp. 95–109.

— (1983), "The geometry of mixture likelihoods: a general theory," *The Annals of Statistics*, 11, 86–94.

Littell, R. and Folks, J. (1971), "Asymptotic optimality of Fisher's method of combining tests," *Journal of the American Statistical Association*, 66, 802–806.

Litton, M. (2009), "Deconvolution in random effects models via normal mixxtures," PhD dissertation, Texas A&M University.

Loughin, T. M. (2004), "A systematic comparison of methods for combining p-values from independent tests," *Computational statistics & data analysis*, 47, 467–485.

Marks, R. J., Wise, G. L., Haldman, G., D., and Whited, J. L. (1978), "Detection in Laplace noise," *IEEE Transactions on Aerospace and Electronic Systems*, 14, 866.

Massey, F. (1950), "A note on the power of a non-parametric test," *The Annals of Mathematical Statistics*, 21, 440–443.

Mudholkar, G. S. and George, E. O. (1977), "The logit statistic for combining probabilities-an overview," Tech. rep., DTIC Document.

Neyman, J. (1937), "'smooth test' for goodness-of-fit," *Skandinavisk Aktuarietidskrift*, 20, 149–199.

Parker, R. and Rothenberg, R. (1988), "Identifying important results from multiple statistical tests," *Statistics in medicine*, 7, 1031–1043.

Pearson, K. (1900), "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *Philosophical Magazine*, 50, 157–175.

Pham, H. (2006), *Springer handbook of engineering statistics*, Springer Science & Business Media.

Quesenberry, C. and Miller, F. L. (1977), "Power studies of some tests for uniformity," *Journal of Statistical Computation and Simulation*, 5, 169–191.

Randles, R. H. and Wolfe, D. A. (1979), *Introduction to the theory of nonparametric statistics*, New York: John Wiley & SONS.

Rayner, J. C. W., Thas, O., and Best, D. J. (2009), *Smooth Tests of Goodmess of Fit Using R*, New York: Wiley, 2nd ed.

Rudzkis, R. and Bakshev, A. (2013), "Goodness of Fit Tests Based on Kernel Density Estimators," *Informatica*, 24, 447–460.

Seshadri, V., Csorgo, M., and Stephens, M. (1969), "Tests for the exponential distribution using Kolmogorov-type statistics," *Journal of the Royal Statistical Society. Series B (Methodological)*, 499–509.

Shapiro, S. S. and Wilk, M. (1972), "An analysis of variance test for the exponential distribution (complete samples)," *Technometrics*, 14, 355–370.

Shapiro, S. S. and Wilk, M. B. (1965), "An Analysis of Variance Test for Normality(Complete Samples)," *Biometrika*, 52, 591–661.

Smith, R. M. and Bain, L. J. (1976), "Correlation type goodness-of-fit statistics with censored sampling," *Communications in Statistics-Theory and Methods*, 5, 119–132.

Song, K. (2002), "Goodness-of-Fit Tests Based on Kullback-Leibler Discrimination Information," *IEEE Transactions on Infromation Theory*, 48, 1103–1117.

Stephens, M. A. (1974), "EDF statistics for goodness-of-fit and some comparisons," *Journal of the American Statistical Association*, 69, 730–737.

— (1978), "On the half-sample method for goodness-of-fit," *Journal of the Royal Statistical Society. Series B*, 64–70.

Sürücü, B. (2008), "A power comparison and simulation study of goodness-of-fit tests," *Computers & Mathematics with Applications*, 56, 1617–1625.

Teicher, H. (1961), "Identifiability of mixtures," *The Annals of Mathematical statistics*, 32, 244–248.

Thompson, R. (1966), "Bias of the one-sample Cramér-Von Mises test," *Journal of the American Statistical Association*, 61, 246–247.

Tiku, M. (1980), "Goodness of fit statistics based on the spacings of complete or censored samples," *Australian Journal of Statistics*, 22, 260–275.

Vasicek, O. (1976), "A Test for Normality Based on Sample Entropy," *Journal of the Royal Statistical Society. Sereis B*, 38, 54–59.

Wang, F. and Keats, J. B. (1995), "Improved percentile estimation for the two-parameter Weibull distribution," *Microelectronics Reliability*, 35, 883–892.

Watson, G. S. (1961), "Goodness-of-fit tests on a circle," *Biometrika*, 48, 109–114.

Yule, G. U. and Kendall, M. (1950), *An introduction to the theory of statistics*, London: Griffin.

Zhan, D. and Hart, J. D. (2012), "Testing equality of a large number of densities," *Biometrika*, 99, 1–17.

APPENDIX A

Theorem A.0.1 *(Randles and Wolfe, 1979, Theorem 5.2.7) Let* $\{S_{n_i}\}$ *and* $\{T_{n_i'}\}$ *be two sequences of tests, with associated sequences of numbers* $\{\mu_{S_{n_i}}(\theta)\}$, $\{\mu_{T_{n_i'}}(\theta)\}$, $\{\sigma^2_{S_{n_i}}(\theta)\}$ *and* $\{\sigma^2_{T_{n_i'}}(\theta)\}$ *and satisfying the following Assumptions A1-A6:*

A1.

$$\frac{S_{n_i} - \mu_{S_{n_i}}(\theta_i)}{\sigma_{S_{n_i}}(\theta_i)} \text{ and } \frac{T_{n_i'} - \mu_{T_{n_i'}}(\theta_i)}{\sigma_{T_{n_i'}}(\theta_i)}$$

have the same continuous limiting $(i \to \infty)$ *distribution with c.d.f.* $H(\cdot)$ *and interval support when* θ_i *is the true value of* θ.

A2. Same assumption as in A1 but with θ_i *replaced by* θ_0 *throughout.*

A3.

$$\lim_{i\to\infty} \frac{\sigma_{S_{n_i}}(\theta_i)}{\sigma_{S_{n_i}}(\theta_0)} = \lim_{i\to\infty} \frac{\sigma_{T_{n_i'}}(\theta_i)}{\sigma_{T_{n_i'}}(\theta_0)} = 1.$$

A4.

$$\frac{d}{d\theta}[\mu_{S_{n_i}}(\theta)] = \mu'_{S_{n_i}}(\theta) \text{ and } \frac{d}{d\theta}[\mu_{T_{n_i'}}(\theta)] = \mu'_{T_{n_i'}}(\theta)$$

are assumed to exist and be continuous in some closed interval about $\theta = \theta_0$ *with* $\mu'_{S_{n_i}}(\theta_0)$ *and* $\mu'_{T_{n_i'}}(\theta_0)$ *both nonzero.*

A5.

$$\lim_{i\to\infty} \frac{\mu'_{S_{n_i}}(\theta_i)}{\mu'_{S_{n_i}}(\theta_0)} = \lim_{i\to\infty} \frac{\mu'_{T_{n_i'}}(\theta_i)}{\mu'_{T_{n_i'}}(\theta_0)} = 1.$$

A6.

$$\lim_{i\to\infty} \frac{\mu'_{S_{n_i}}(\theta_0)}{\sqrt{n\sigma^2_{S_{n_i}}(\theta_0)}} = K_S \text{ and } \lim_{i\to\infty} \frac{\mu'_{T_{n_i'}}(\theta_0)}{\sqrt{n'\sigma^2_{T_{n_i'}}(\theta_0)}} = K_T$$

where K_S and K_T are positive constants. Then

$$ARE(S,T) = \frac{K_S^2}{K_T^2}.$$